機器學習書面報告

主題 ： 資料科學相關行業薪資分析

組員: 阮柏誠、薄育文

摘要

本研究使用的資料集為 2020 到 2023 的數據科學家薪資，由 aijobs.net 網站上蒐集 ，並於 2024/11/02 於 kaggle 下載。內容主要分為兩部分，第一部分為對於資料的探索式 分析，第二部分為針對預測薪水高低進行建立機器學習的相關模型，以預測並探討何種 因素成就數據科學家薪水的高低薪資。

第一章 緒論

第一節 研究動機

隨著科技快速發展，大數據已成為商業決策的重要基礎，資料分析師等相關職位 也變得不可或缺，我們想要了解資料相關領域裡有哪些潛在的職業機會。例如，這些工 作的職稱是什麼？需要什麼樣的技能？未來發展前景(薪資)怎麼樣？了解這些能幫助 我們更明確地規劃努力的方向，也可以為未來的學習和職涯設定更清晰的目標。而且， 隨著科技的蓬勃發展，數據在各行各業已經成為商業決策的關鍵資產，資料分析相關的 工作在未來一定會越來越重要。

第二節 研究目的

我們想透過機器學習的方法分析哪些工作職稱傾向較高的薪資。而薪水 不僅代表市場需求，也和這些工作需要的專業程度、挑戰性有很大的關係。如果能搞清 楚高薪職位的特點，我就可以更有針對性地學習、提升自己的能力，讓自己在未來的職 場上更有競爭力，也能更快實現自己的職業和個人價值。 所以我們希望能全面了解資料科學相關行業的職業機會，特別是那些既有發展 潛力又能帶來高收入的工作，這樣才能找到最適合自己的發展方向。

第三節 研究流程

第一章將詳述本研究的動機與目的，說明研究背景與期望達成的目標。第二章則 介紹資料來源，探討資料中的問題及其處理方式，並進行資料的探索性分析，以挖掘資 料中的關鍵資訊。 第三章將著重於工作職稱變數的分群，並在變數不足的情況下，透過拆解職稱字 詞來增加特徵欄位。同時，我們希望分析不同字詞對薪資高低的影響。第四章的目標是 建構模型來預測薪資水平。我們將運用四種模型進行薪資預測，並比較其效能，解釋哪 些特徵和因素對薪資的高低產生影響。 最後，第五章將總結研究過程中的關鍵成果與發現，並提出未來研究的建議與方向。

第二章 資料介紹與探討

第一節 變數介紹

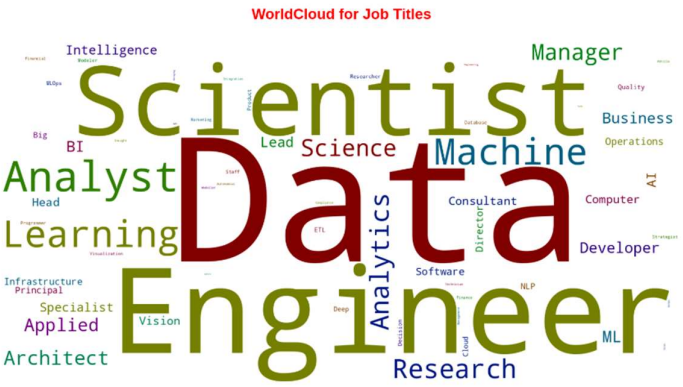
變數一共有 11 個，但扣除重複性值的剩餘以下 8 個，分別有職稱、雇用類型、經驗水平、公司所在地、薪資、員工居住地、公司規模與年份，詳細介紹請見表一

變數	摘要
Job Title	職稱，共 111 個職位名稱，如資料科學家、資料分析師
Employment Type	雇用類型(Full-Time、Contract、Part-Time、Freelance)
Experience Level	經驗水平(Entry、Mid 、Senior 和 Executive)
Company Location	公司所在地點，共有 71 個國家，如美國、英國等
Salary in USD	薪資，以美元紀錄，最少 15000 美元，最多 450000 美元
Employee Residence	員工居住地，共有 83 個國家，如美國、英國等
Company Size	公司的規模，分為大、中、小
Year	年份，時間從 2020 年到 2024 年

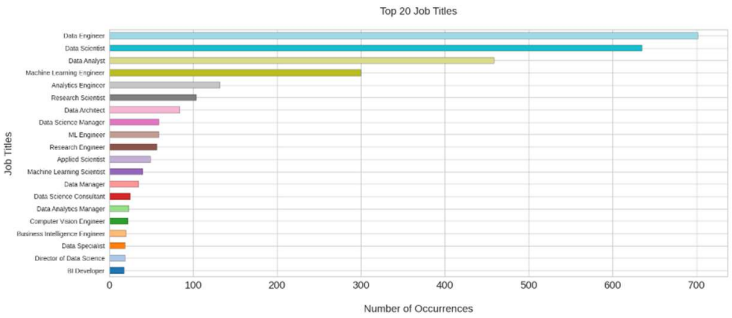
表一 變數介紹

第二節 EDA

在這小節中，我們將透過一些圖形來幫助我們初步的認識資料，首先由於我們想 了解工作職稱的種類及其中哪些字詞較為重要，根據(圖一)我們可以得知，職稱中重複 出現最多的字詞是 Data，其次是 Engineer、Scientist 等。透過圖二，我們發現 Data Engineer、Data Science、Data Analysis 與 Machine Learning 等職業的數量遠超過其他 職業。然而，根據圖三，雖然這些職業的數量在圖二中占比較大，但其餘職業的總和仍 占 36.5%。因此，接下來將依照不同方式對工作職稱進行分群。



圖一、 Job Title 文字雲

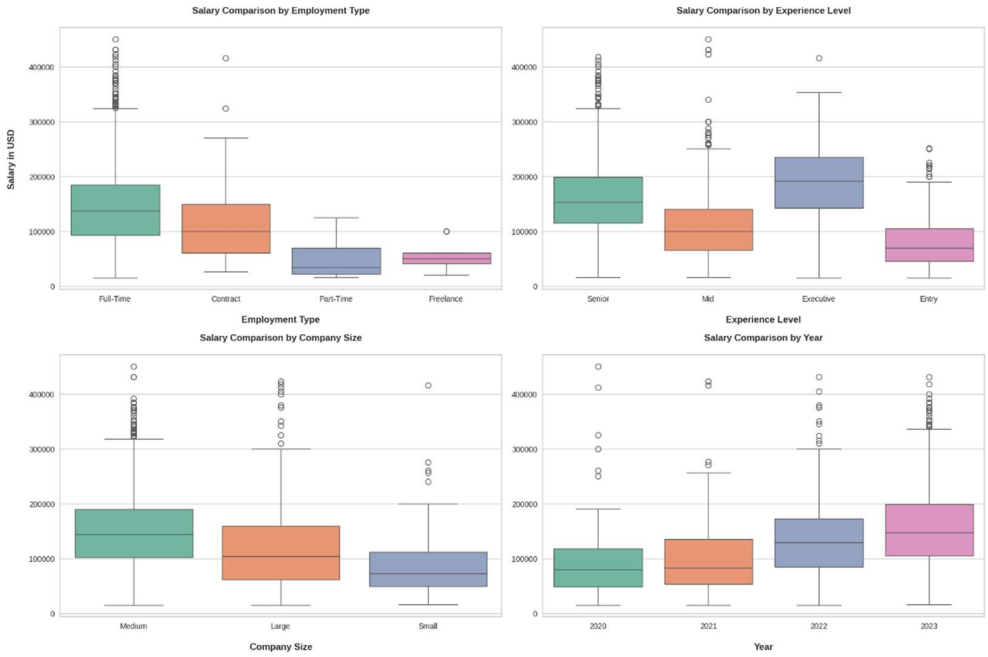


圖二、 Top 20 Job Title

接下來是關於公司所在地與員工居住地圓餅圖從圖三中，我們可以得知無論是公司所在地亦或是員工居住地，美國的比例都約若在 75%，且英國所在地皆為次之大約 7.5%，至於其他所在地約為 17%到 18%左右。且發現公司所在與員工居住地的比例有些微差異，原因之一可能是遠距工作導致員工居住地與公司所在地可能會有些許不一致。

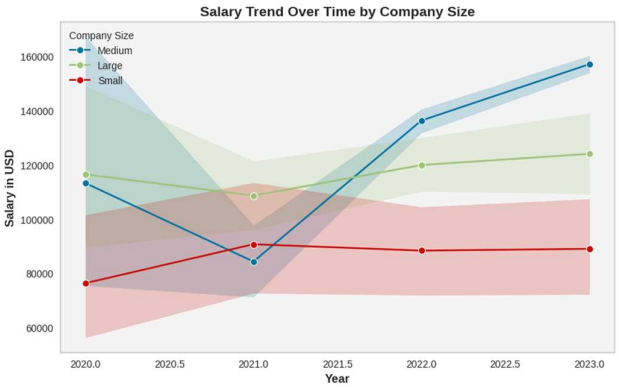
再來我們想要知道對於薪水，雇用類型、經驗水平、公司的規模與年份是否有明顯的差異，因此我們將這些類別對薪水畫盒鬚圖，來判斷是否有差異。結果如圖四所示，確實不同的雇用類別、經驗水平、公司規模與年份各自皆對薪水有差異，不過，令人感到意外的是在公司規模與薪資的部分以中型公司給予的薪資較多，至於原因如何，我們可以再透過圖五來判別。

圖三 公司所在地與員工居住地圓餅圖

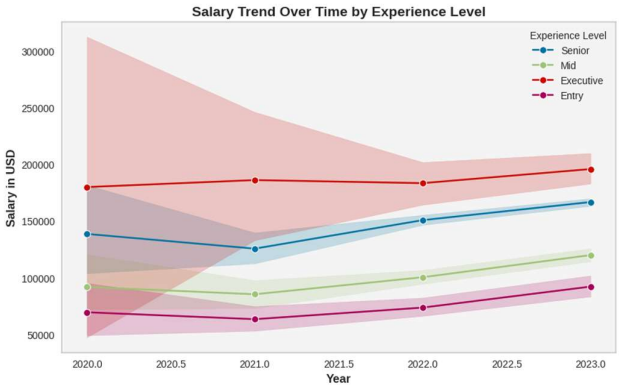


圖四 依照雇用類別、經驗水平、公司規模與年份對薪水畫盒鬚圖

圖五為不同公司規模的薪資與時間變化，透過此圖，我們可以發現，中等公司在 2021 年到 2023 年裡，給予數據科學家的薪資 逐年上升，甚至高過於大型公司，我們猜測是因為近年來數據科學家越來越受到重視，若要讓公司在近一步，就可能會很需要數據科學家的幫助，因此給予的薪資有機會超過大公司給予的穩定薪資。而透過圖六我們可以知道經驗對於薪水的影響是有的且不會因為受到時間的因素導致薪水有較高等的經驗被較低經驗超越的問題。



圖五 不同公司規模的薪資隨時間變化

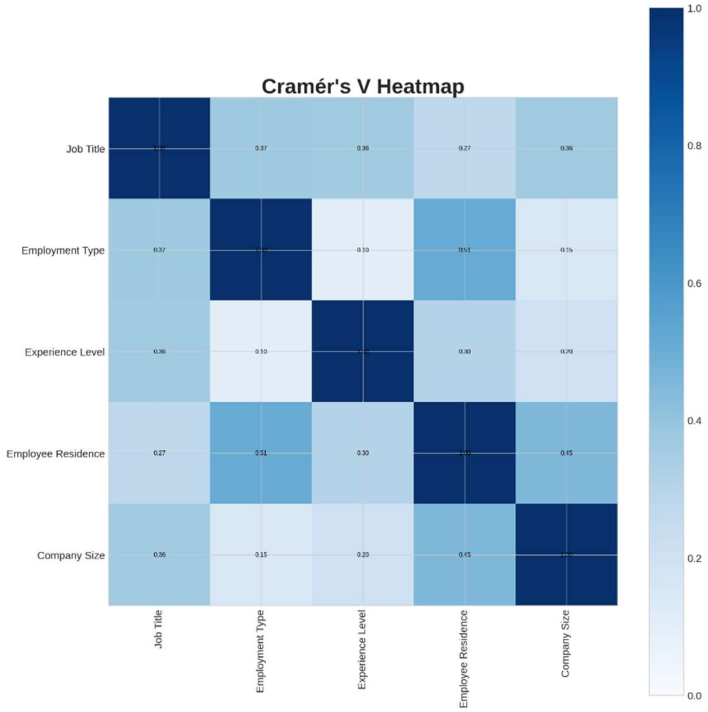


圖六 不同經驗水平的薪資隨時間變化

最後透過Cramer's V 來對類別型變數畫相關係數圖。Cramer's V 是一種用來衡量兩個類別變 數之間關聯強度的統計量

$$\textbf{Cramer's V 的計算公式：} \quad V = \sqrt{\frac{\chi^2}{n \times (k-1)}}$$

其中 χ2 為卡方統計量，n 是樣本的總數，k 是行數或列數中較小的數，透過此公式在畫出 來的相關係數圖如圖七所示，我們可以發現工作職稱與其餘變數皆有約 0.3 的相關性 ，公司大小對於年份以及員工居住地也有約 0.3 左右的相關性，而最有趣的是雇用類型與員工居住地有約 0.7 的相關性，不過這可能有一部份是因為雇用類型以全職工占了 99%，而員工居住地以美國占了 75%所導致的。



圖七、 類別相關係數圖

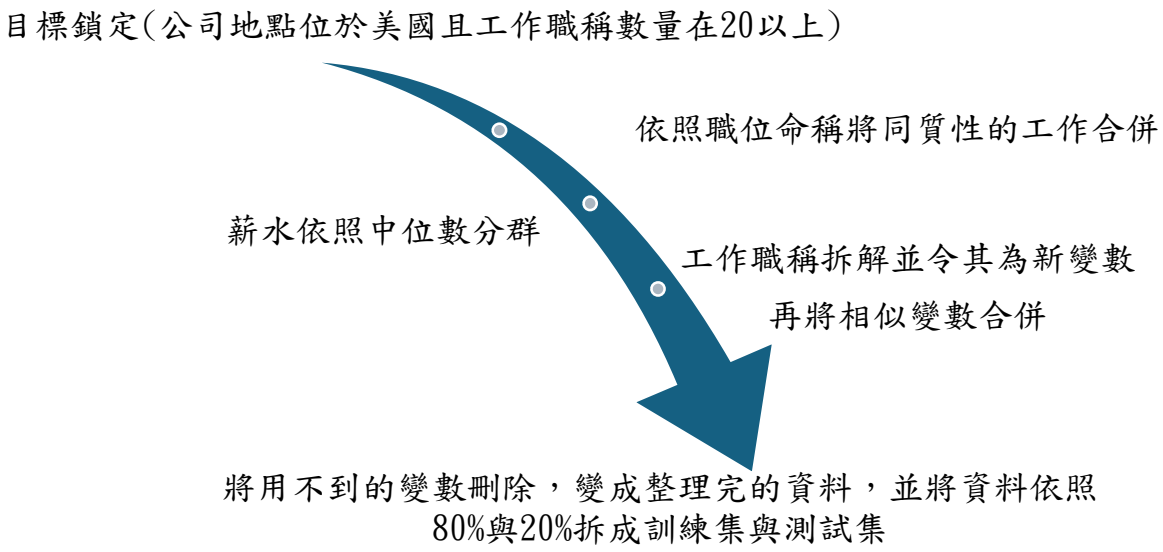
第三節 小結

在這章節中，我們初步的認識資料，由於許多變數資料占比不均，所以後續我們的目標會著重於公司所在地位於美國且工作名稱至少出現 20 筆以上，並對薪水藉由中位數分兩群的前提下進行後續的探討以及建模。

第三章 資料處理與資料探勘

在上一章節中，我們發現許多變數資料占比不均，因此我們將聚焦在位於美國的公司，且由於工作名稱非常多種因此我們將目標鎖定在出現 20 筆以上的資料，並且將同質性高的工作合併，如 Machine Learning Engineer 與 ML Engineer 合併為新的 Machine Learning Engineer。再來我們將對薪水依照中位數做分組，因為我們的變數全都是類別變數，鎖定在何種原因造就資料科學家的高薪與低薪。

我們猜測或許不同的工作名稱的字詞對薪水或許有一定的影響，為此我們對職稱的字詞做拆解，先令其為新的變數再將相似字詞合併，若該職稱有包含該字詞，則該變數欄位標記為 1。令其為新變數則如 Data Science，在 Data 與 Science 欄位記為 1，其餘字詞欄位記為 0。而相似字詞合併如 Analyst 與 Analytics 我們會合併成 Analyst，Scientist 與 Science 會合併為 Scientist 等。將上述步驟做完後我們將對整理完成的資料依照訓練集與測試集並拆成 8 比 2，以利後續帶入後續模型。



圖八 資料處理與資料探勘的流程圖

第四章 模型建立和模型比較

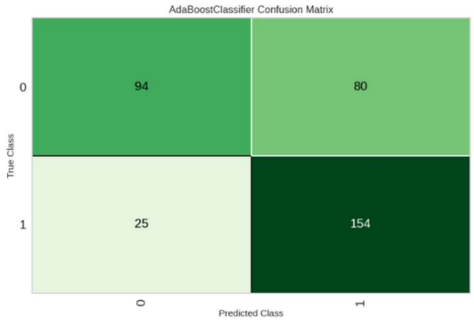
再這章節中，我們將透過建立三個機器學習的方法與一個深度學習中最基礎的 DNN 來建立我們的模型，其中機器學習分法使用了 Ada Boost Classifier、Light Gradient Boosting Machine 與 XGBoosting Model 的模型。其中機器學習的模型我們將使用 pycaret 套件來建立，每個機器學習模型，對類別變數的種類若小於等於 10 做 one hot encoding，大於 10 則做 target encoding，並透過 3 折交叉驗證來讓模型更穩定，我們的預測目標為我們的薪水分群(0 代表小於中位數 1 代表大於等於中位數)。

一、 Ada Boost Classifier Mode

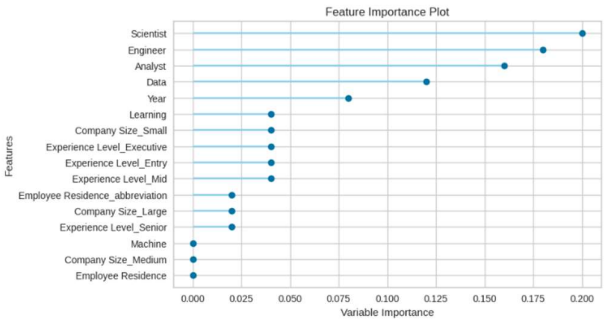
本模型使用 AdaBoost 分類器，並將訓練結果應用於驗證集，其混淆矩陣如圖八所示。混淆矩陣顯示，類別 0（低薪）的分類錯誤較多，表示模型對低薪族群的預測誤差較高。

透過圖九，我們可以觀察模型中不同變數的重要性。其中，職稱拆解後的字詞中有四個較為重要，分別為 Scientist、Engineer、Analyst 和 Data，而 Year 的重要性則次之。

最終預測結果如表二所示，模型在驗證集與測試集上的表現約為 0.6 多。其中，AUC 在驗證集中達 0.72，但在測試集僅 0.46，顯示模型在測試集上的區分能力較低。此外，F1-score 在驗證集中為 0.72，而在測試集中降至 0.62，反映出模型的泛化能力仍有待提升。



圖八、Ada Boost 混淆矩陣



圖九、Ada Boost 特徵重要性

	Accuracy	Auc	F1-score
驗證集	0.67	0.72	0.72
測試集	0.63	0.46	0.62

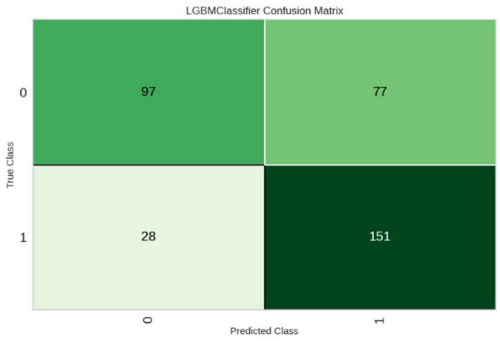
表二 Ada Boost 預測結果

二、 Light Gradient Boosting Machine

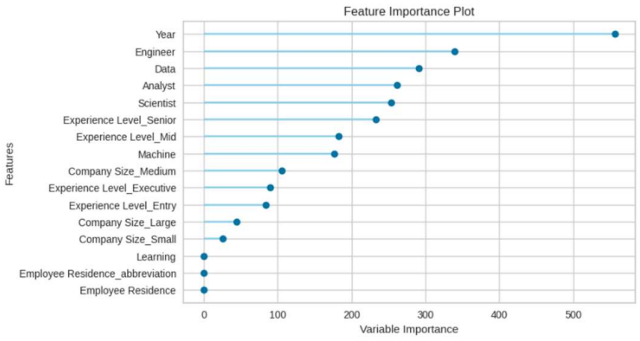
本模型使用 Light Gradient Boosting Machine 分類器，並將訓練結果應用於驗證集，其混淆矩陣如圖十所示。混淆矩陣顯示，類別 0（低薪）的分類錯誤較多，表示模型對低薪族群的預測誤差較高。透過圖十一，我們可以觀察模型中不同變數的重要性。其中該模型以 Year(年份)較為重要，Scientist、Engineer、Analyst 和 Data 次之，而經驗水準重要程度稍為弱於職稱拆解後的四個字詞。

接下來，透過可解釋 AI 的方法進一步分析模型訓練結果，結果如圖十四所示。在圖中，紅色出現在左側表示該變數會使薪水趨向低薪區間，而紅色出現在右側則表示該變數會使薪水趨向高薪。變數的重要程度依照上到下排序，其中 Data 和 Analysis 影響薪水往低薪區間發展，而經驗水準較高或具專業技能者則更可能獲得較高薪資。另一方面，缺乏經驗的求職者較容易落入低薪範疇，而 Scientist 這一職稱則與高薪較為相關，顯示出該職稱在薪資上的影響力較大。所以我們知道以該模型而言，我們想要往高薪走，應該要把自己的經驗提高，且在搜尋數據科學家工作時，應該要往 Scientist 這個字眼找起。

最終預測結果如表三所示，模型在驗證集與測試集上的表現約為 0.6 多。其中，AUC 在驗證集中達 0.69，但在測試集僅 0.44，顯示模型在測試集上的區分能力較低。此外，F1-score 在驗證集中為 0.70，而在測試集中降至 0.62，反映出模型的泛化能力仍有待提升



圖十 LGB 混淆矩陣



圖十一 LGB 特徵重性

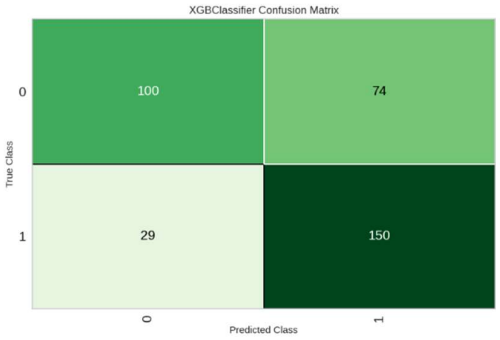
	Accuracy	Auc	F1-score
驗證集	0.66	0.69	0.70
測試集	0.63	0.44	0.62

表三 Light Gradient Boosting Machine 結果

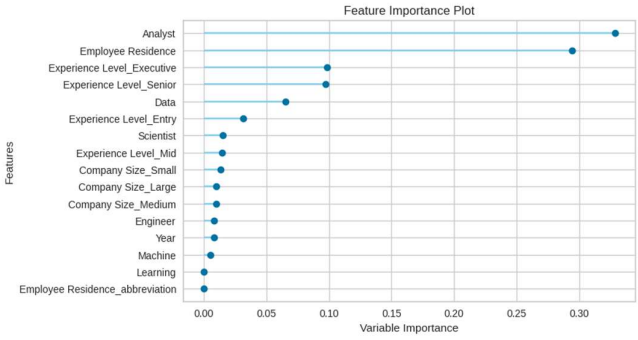
三、 XGBoosting Model

本模型使用 XGBoosting Model 分類器，並將訓練結果應用於驗證集，其混淆矩陣如圖十二所示。混淆矩陣顯示，類別 0（低薪）的分類錯誤較多，表示模型對低薪族群的預測誤差較高。透過圖十三，我們可以觀察模型中不同變數的重要性。其中該模型以 Analyst 與 Employee Residence 較為重要，次之重要為專業與高等的經驗水準。而透過可解釋 AI 的方法進一步分析模型訓練結果，結果如圖十五所示。其結果與 LGB 的 shape 差不多，若我們想要往高薪走，應該要把自己的經驗提高，且在搜尋數據科學家工作時，應該要往 Scientist 這個字眼找起。

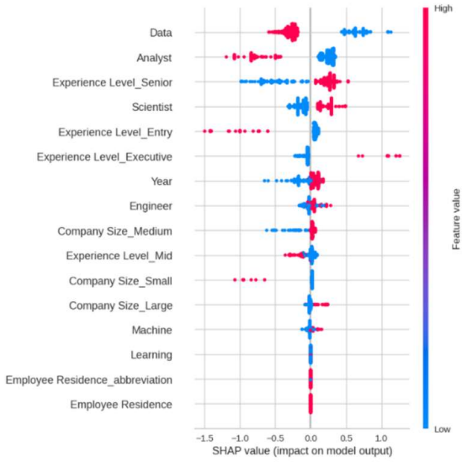
最終預測結果如表三所示，模型在驗證集與測試集上的表現約為 0.6 多。其中，AUC 在驗證集中達 0.69，但在測試集僅 0.43，顯示模型在測試集上的區分能力較低。此外，F1-score 在驗證集中為 0.70，而在測試集中降至 0.62，反映出模型的泛化能力仍有待提升



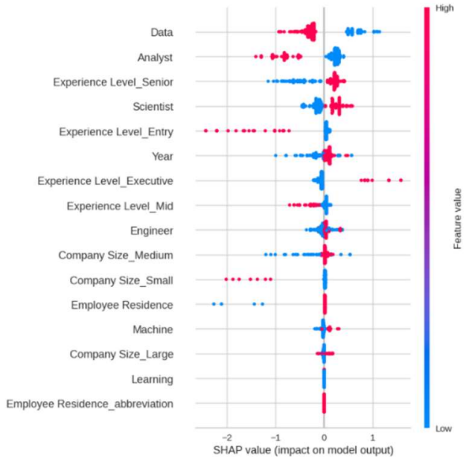
圖十二 XGBoost 混淆矩陣



圖十三 XGBoost 特徵重性



圖十四 LGB shap 圖



圖十五 XGBoost shap 圖

	Accuracy	Auc	F1-score
驗證集	0.65	0.68	0.70
測試集	0.63	0.43	0.62

表四 XGBoosting Model 結果

四、DNN.

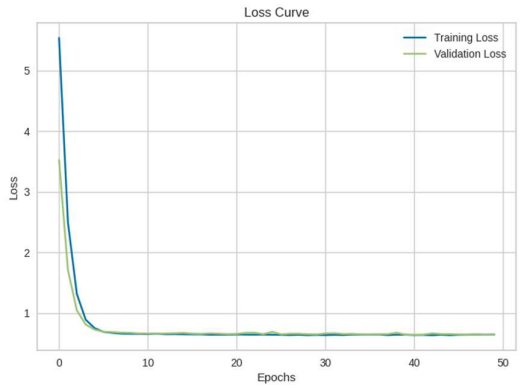
由於我們透過機器學習的方法的成效 Accuracy 大約都若在 0.65 左右，所以我們猜測可能透過機器學習的方法無法在更好，因此我們透過深度學習的方式來建模，希望能提高預測能力。圖十六為我們的神經網路架構與神經元個數，而我們的 epoch 為 100，圖十七是我們學習取線，可以發現該模型的 Loss Function training loss 與 vaildatin loss 一致，代表訓練效果已經達到上限，而 DNN 的結果如表五所示測試集為 Accuracy 為 0.71，Auc 為 0.75，F1-score0.74，結果都優於其他機器學習的方法。

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 512)	8,704
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 218)	111,834
dropout_1 (Dropout)	(None, 218)	0
dense_2 (Dense)	(None, 218)	47,742
dropout_2 (Dropout)	(None, 218)	0
dense_3 (Dense)	(None, 128)	28,032
dropout_3 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 1)	129

Total params: 196,441 (767.35 KB)
Trainable params: 196,441 (767.35 KB)
Non-trainable params: 0 (0.00 B)

圖十六 DNN 參數



圖十七 DNN 學習曲線

	Accuracy	Auc	F1-score
測試集	0.71	0.75	0.74

表五 DNN 結果

五、模型比較：

透過建立四個模型來做預測，機器學習方法效果的上限無論用何種指標大約都落在 0.65 到 0.7 附近，只有在測試集的 AUC 會在 0.4 多，而深度學習各種指標都若在 0.7 以上，效果也都比機器學習的方式好，因此若要選個模型的話，會選擇 DNN 為我們的模型。

第五章 結論

本研究針對 2020 至 2023 年間數據科學家的薪資數據進行了探索性分析與機器學習建模，旨在探討影響薪資高低的關鍵因素。透過資料分析，我們發現職稱、經驗水平、公司所在地、雇用類型 及 公司規模 等因素對薪資水準具有顯著影響。其中，職稱中的關鍵詞如 Scientist、Engineer、Analyst 等對薪資高低扮演重要角色，而 隨著經驗水平的提升，薪資亦呈現正向成長。

在機器學習建模方面，我們採用了 AdaBoost、LightGBM、XGBoost 及 DNN 等模型來預測薪資水準，並透過交叉驗證評估模型表現。結果顯示，DNN 在各項指標中表現最佳。然而，若希望進一步提升模型準確性，可以考慮 蒐集更多數據，並採用更豐富的特徵工程方法來挖掘潛在資訊。

此外，透過可解釋 AI 技術，我們深入探討了模型對不同變數的影響，結果發現 職稱包含 "Scientist" 的職位與高薪較為相關，而 "Data" 與 "Analysis" 等詞彙則較常出現在低薪職位中。此發現可為未來求職者提供參考，若希望提升薪資水準，應強化自身技能與經驗，並選擇更具競爭力的職稱與職務。

若針對高薪職缺的尋求，本研究提出以下建議：優先聚焦於 Scientist 相關職業，因這些領域的薪資水準普遍較高，而別聚焦在 Analysis。持續累積專業經驗與提升技能水平，以增強競爭力，為實現高薪目標奠定堅實基礎。透過本研究的分析與建議，希望能為數據科學領域的從業者與求職者提供有價值的參考，幫助其在職涯發展上做出更明智的選擇。

參考資料

資料來源：

https://www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023