模型訓練及選用問題　　評測者　商研院　朱靖文

問題

1. 如果要達成「能夠預測指定的symbol在90天後，是否有成長10%」的目標，會選用的模型及訓練方式，文字提供選用的模型及原因

根據所知所學，以機器學習的模型來說，要能預測指定的symbol在90天後成長「多少」？可以使用線性迴歸的模型來解決迴歸問題；而要能預測指定的symbol在90天後「是否」成長？則可以使用邏輯迴歸、決策樹、KNN等模型來解決分類問題。

若考量僅是訓練模型判斷是否成長10%可以透過現有資料確立每90天（每一季）的成長比例作為標籤欄位，並將標籤處理分為成長大於等於10%及成長小於10%兩種，隨後以預測分類標籤之方式，經由篩選過的特徵欄位判斷是否得到符合預測之標籤欄位，從而驗證模型之準確率，此方法解決分類問題之模式，主要是針對問題結果所做之目標導向。

若考慮將過程中之相關性、迴歸係數以及自變數和應變數之關係，則可選用線性迴歸模型，首先，可根據特徵欄位需求多寡決定簡單線性迴歸或多元線性迴歸，接著，透過現有資料確立每90天（每一季）的成長比例之數值，藉由此數值以便訓練模型預測得到預期成長率之數值，最後，若需要「是否」的標籤，則可以另外製作轉換標籤的條件判斷，同樣可以透過模型得出預測結果，且可經由標籤轉換回推預測結果之成長比例可能的數值，對於未來決策的檢討反思及驗證核對有更多要素能參照。

2. 在訓練中如何從現有的資料集提取出關鍵影響欄位

為判斷何謂有關鍵影響之欄位，根據統計知識，會先作相關性分析（Correlation Analysis）對每一個特徵欄位進行熱力圖分析相關係數之數值，並將具有高度相關（值介於0.6與0.9之間）之特徵欄位提取進行判斷、再計算、選擇。倘若相關性得出的結果差不多高低水平，則需要再進一步使用判斷特徵重要性的技術，譬如隨機樹、相互資訊（Mutual information）、遞迴特徵刪除（RFE）等等。甚至是統計學中之ANOVA F-test作為解決分類問題之模型的方法。

3. 如何利用目前已有的資料集欄位，推論出更有效的新資料欄位

除了對於現有的資料欄位有一定程度的理解之外，將現有的資料欄位進行計算是一個方法，對單一欄位的數據進行處理可對該特徵的總體數據進行壓縮或放大，譬如開根號或平方，又或者依照比例換算為百分比；對於多個欄位亦可以進行數據的變換及處理，透過彼此相加、相減、相乘、相除之後建立新的特徵欄位，譬如open和close進行相加後的特徵欄位、high和low計算差值的特徵欄位、change和changeOverTime得出單位時間變化的比值。