

ML_final_report__第N組



主題 - Machine Learning Report for Predicting the Success of Startup Companies

在當今競爭激烈的商業環境中，新創公司的成功與否對於投資者、創業者以及風險投資基金等利益相關者來說至關重要。然而，傳統的評估方法往往基於經驗判斷和主觀評估，存在主觀性、局限性和風險高的問題。為了克服這些挑戰，機器學習技術的快速發展為我們提供了一種新的途徑。

本報告的主題是利用機器學習預測新創公司成功的潛力。通過利用大量的歷史數據和強大的計算能力，我們可以建立預測模型，從中學習模式並進行預測。這樣的方法可以幫助投資者和創業者更準確地評估新創公司的成功潛力，從而作出更明智的投資和創業決策。我們使用了六種不同的機器學習分類器模型，包括Logistic Regression、Decision Tree Classifier、Random Forest Classifier、GaussianNB、Gradient Boosting Classifier和AdaBoost Classifier，並透過Grid Search CV方法調優這些模型的參數。

最終，我們的目標是推動機器學習技術在投資和創業領域的應用，展示其在預測和決策支持方面的潛力。通過實際應用和驗證，我們將為投資者、創業者和風險投資基金等利益相關者提供可靠的參考和指導。



動機

預測新創公司的成功潛力並非易事，因為它涉及到諸多變數和不確定因素。機器學習技術的發展提供了一種新的途徑，可以利用過去的數據和模式來預測未來的結果。

本報告的主要動機在於：

1. 提供一種基於機器學習的方法，可以幫助投資者和創業者更準確地評估新創公司的成功潛力，從而作出更明智的投資和創業決策。
2. 增進對新創公司成功與否影響因素的理解，通過分析大量的歷史數據，發現潛在的關聯性和趨勢，從而提供洞察和預測。
3. 推動機器學習技術在投資和創業領域的應用，展示其在預測和決策支持方面的潛力。

這些動機將有助於促進新創公司生態系統的健康發展，減少風險並提高成功率。同時，對於投資者和創業者來說，這項研究可以提供更多可靠的參考和指導，使他們能夠做出更明智的投資和創業決策，從而推動創新和經濟發展。



報告貢獻性

新創公司的成功潛力一直是企業界關注的焦點。本報告旨在利用機器學習技術，基於可用數據對新創公司的成功與否進行預測。我們將介紹使用的數據集、所選機器學習模型、特徵選擇以及模型評估的結果。模型基於大量的歷史數據進行訓練，並能夠根據新創公司的特徵預測其未來的成功概率。這項研究有助於投資者、創業者和風險投資基金等利益相關者做出更明智的決策。



（一）數據集介紹：

- 數據集選擇與來源：

我們的資料來源主要是尋找kaggle所提供的數據集。原先我們決定選擇一個數據集，並將其切割為訓練數據集與測試數據集，就可以利用測試數據集來做預測。

但因為以下原因，我們決定使用兩個不同的數據集：

1. 驗證模型的泛化能力：

當我們使用一個不同的數據集來進行測試時，我們可以評估模型在未見過的數據上的表現。這可以提供對模型如何處理新的、未知數據的實際效果的理解。

2. 減少overfitting的風險：

使用不同的數據集來進行測試可以幫助我們確定模型是否發生了overfitting。如果模型在訓練數據上表現很好，但在測試數據上表現不佳，這可能是overfitting所導致。透過使用其他數據集進行測試，有效減少overfitting的風險。

3. 驗證模型的可靠性和穩定性：

使用不同的數據集來進行測試可以驗證模型的可靠性和穩定性。如果模型在不同的數據集上都能表現良好，那麼我們可以更有信心地相信模型的效果是可靠的。

綜合上述原因，我們使用了兩個數據集。除了可以對原先的數據集進行訓練與測試外，可以再測試另一數據集，觀察模型在不同數據集的預測表現。

而我們選擇數據集所考慮的特點如下：

1. 特徵數量多：

較多的特徵數量可以提供更多的信息和細節，使模型能夠更好地捕捉數據中的模式和關聯性。同時，不同類別之間的差異和變異性也更容易被模型捕捉到，提高模型的預測能力和泛化能力。

2. 多樣性:

包含較多不同種類的公司與多樣的情況，可以使模型更全面地學習到不同類別之間的差異和變異性。這樣的數據集可以幫助模型捕捉到更多的細節、趨勢和特徵，從而提高模型的性能和預測能力。

3. 大規模:

大規模數據集通常包含更多的樣本和觀測值，從而提供了更全面的數據覆蓋。同時也能提供更穩定和可靠的統計性能。隨著數據量的增加，統計推斷和估計的可信度也會增加，能夠減少抽樣偏差和隨機噪音的影響，提供更準確的結果。

4. 相似性:

如果兩個數據集的差異過大，模型可能無法有效地利用測試數據中的特徵進行預測，在測試數據上的性能可能與在訓練數據上的表現差距很大。因此訓練數據集和測試數據集應具有一定的相似性，確保模型在未見過的數據上能夠有效地泛化。

依照上述特點，我們選擇了兩個數據集，因為一個規模較大簡稱為big，另一數據集簡稱為small。我們利用big訓練出模型，再利用small進行預測，觀察模型的泛化能力。

兩者的資料來源如下:

▼ big的數據集連結: <https://www.kaggle.com/datasets/yanmaksibig-startup-secsees-fail-dataset-from-crunchbase/code>

▼ small的數據集連結: <https://www.kaggle.com/datasets/manishkc06/startup-success-prediction>

• 數據預處理：

數據預處理主要分為數據清理、特徵篩選、新增特徵與資料轉換四個部分。

數據清理

將缺失值與重複數據刪除，並將沒有定義清楚的特徵刪除。其次是去除一些不合理的數值，例如刪除在時間相關變數中出現小於0的資料（資料集中的數字應該大於等於0）。最後，因數據集規模較大，資料太過繁瑣，所以選擇留下存在較多數據的城市，並將其他數據較零碎的城市刪除，避免模型在稀有城市上的overfitting。

特徵篩選

篩選出最終要使用的特徵，下一個小節中會更詳細的介紹我們如何在剩下的特徵中挑選出我們認為有解釋力的特徵值與變數。

新增特徵

新增我們需要的特徵。例如資料中原本只有新創公司成立的日期，我們希望能以月份與年份搭配季度的維度討論時間對新創成功的影響，因此新增新創公司創立的月份與年份季度兩個特徵。

資料轉換

首先是新創種類的重新分配，在透過兩個資料集比對之後，透過數量與重疊性選擇好目標的產業種類，並依據相關的關鍵詞將新創公司重新分配到我們訂好的新分類下。接著，將兩資料集的特徵名字與各特徵下的變數類型統一。最後將各特徵的表示方式轉換為我們希望的樣子，例如將第一及最後一筆募資資金流入時公司的年齡以四捨誤入的方式改以整數的方式呈現。

- 特徵提取和選擇：

對新創公司來說決定是否可以繼續經營擴大的主要因素有兩個：產品與金流。產品需要有亮點有發展性才會吸引更多投資者投入資金，公司才有機會利用這些金流做更多或規模更大的產品。然而，產品的市場性與發展性較不容易被量化，加上對於新創公司來說資金流入量可以間接代表產品在投資人眼中的價值，因此我們會以資金流動相關的指標作為主要的特徵值。產業種類雖然不是量化的特徵值，但是我們認為產業不同的產品有較大的異質性，因此資料中有將新創公司分為主要常見的18個產業也作為後續分析時可使用的特徵值。最後是宏觀中可能會受社會的景氣、國家經濟狀況、是否有重大事件等大環境因素影響，導致特定時期或時間對新創發展有整體較活躍或較衰退的影響。

綜合以上提到的影響因素我們主要選擇了三種類型的特徵值：資金相關的特徵值、產業種類、時序相關的特徵值，並在下方做更詳細的介紹。

資金相關的特徵值

特徵值包含募資階段數 (*Funding_rounds*)、第一及最後一筆募資資金流入時公司的年齡 (*Age_first_funding_year / Age_last_funding_year*)、總募資金額 (*Funding_total_usd*)、有無天使投資人 (*Has_angel*)、有無創業投資人 (*Has_VC*)、募資週期中四輪分別有無成功募資 (*Has_roundA, B, C, D*)。

1. 募資階段數：

通常階段數越多代表不斷的有投資者對公司的發展認可，隨著公司成長也願意繼續投入資金，因此預期會有較高成功機會。

2. 第一及最後一筆募資資金流入時公司的年齡：

通常較謹慎的投資者會評估新創公司的產品是否有成熟的發展藍圖、是否有一定程度的穩定合作夥伴或是利益關係人、是否具備一定市場價值等，在滿足這些條件後認定是有機會成功時才會投入資金。第0年未見任何成長就有第一筆資金流入的新創公司不一定代表前景較被看好。因此，將此變數也作為特徵值，看是否成功的新創通常第一和最後一次被投資的時機是有規律的。

3. 總募資金額：

如前述，投資者對新創公司的價值評估會間接反應在投資的金額中，我們會預期較被看好的新創公司具有較高的募資總額。

4. 有無天使投資人：

天使投資人通常投入的資金額不大，通常是提供人脈資源、社會經驗、商業知識輔助的角色。因此有天使投資人確實可以預期新創有較高的成功機會，但仍有很大的風險，可能較具有創業投資人的預測力低。

5. 有無創業投資人：

通常創業投資人是集結許多專業投資人的資金投資新創公司，除了在決定給予投資金時就已經有更高、更專業、更加具有指標性的標準外，也可能會有更大程度的直接進入公司管理、指導與監督公司的發展方向與成長指標。也就是說預期有創業投資人進入可能具有較高的成功機會。

6. 募資週期四輪是否有成功募資：通常越到後期公司發展逐漸穩定風險相對初期較穩定，資金額也會隨著階段數增加而增加。資金如果可以在後期流入，金額較大通常最後成功的機率也較高。

產業種類

觀察兩個資料集中新創在各種類的數量分布，選擇出同時都有較多新創公司的種類，並將部分相似的種類名稱重新分組組成較大的分類。由於兩個資料集的用詞與提供的形式不同，我們決定直接以產業做為劃分方式，重新分類資料集中的公司，並排除數量較少的產業的新創資料，不列為分析用的資料。

最後我們總共列出18個產業種類：

1. web: 包含network hosting, web hosting, web design...
2. games
3. hardware & software: 包含messaging
4. biotechnology
5. analytics
6. e-commerce

7. mobile
8. advertising
9. semiconductors
10. security
11. clean technology
12. public relations
13. social media: 包含social, news, video相關的
14. fashion
15. travel
16. art & music
17. health & medical 18. finance: 包含real estate

以科技新創產業發展的歷史作為產業種類可能會受宏觀市場影響的例子。在1990年代後期開始因為受半導體發展帶動軟硬體的技术高速成長，也促使網絡產業的發展興起，許多新創公司開始趁著熱潮出現。初期許多公司也確實有非常出色的表現（例如：dot-com，在當時就非常成功），許多創業投資人也關注到這個領域的公司，並投入許多資金，但也因為突然性的大家都擠進這個市場導致泡沫效應，後期許多新創公司在泡沫破裂之後都宣告失敗。

時序相關的特徵值

時序部份我們選擇新創公司創立的時間（*Founded_at*）為特徵值並另外創立月份（*Founded_month*）與年度搭配季度（*Founded_quarter*）的變數，雖然這些變數無法用來預測時序差距較大的資料，但希望可以由此分析看出哪一種時間維度較能說明環境對新創成功的預測力。此次的報告中因為是以年度相近的資料集做預測，因此直接將時序變數視為一般可以直接解釋不受時序影響的特徵值使用。



（二）模型選擇：

我們使用了六種不同的機器學習分類器模型，包括Logistic Regression、Decision Tree Classifier、Random Forest Classifier、GaussianNB、Gradient Boosting Classifier和AdaBoost Classifier：

1. LogisticRegression:

- 優點：適用於二元分類問題，具有解釋性好的優勢，對於特徵之間的關係和重要性有較好的解釋能力。
- 缺點：在處理大量特徵和非線性關係時，性能可能受到限制。

```
clf = LogisticRegression()
    parameters = {
        'penalty': ['l1', 'l2'],
        'C': [0.01, 0.1, 1.0, 10.0],
        'solver': ['liblinear', 'lbfgs'],
        'max_iter': [100, 200, 500]
    }
```

2. DecisionTreeClassifier:

- 優點：能夠處理非線性特徵和特徵交互作用，且易於理解和解釋。
- 缺點：容易出現過擬合，特別是在深度過深的樹中，對於噪聲和不確定性較敏感。

```
clf = tree.DecisionTreeClassifier()
    parameters = {
        'criterion': ['gini', 'entropy'],
        'max_depth': [None, 5, 10, 15],
        'min_samples_split': [2, 5, 10],
        'min_samples_leaf': [1, 2, 3]
    }
```

3. RandomForestClassifier:

- 優點：由多個決策樹組成，有效地降低了過擬合的風險，並且在處理大量特徵時具有良好的性能。
- 缺點：可能難以解釋單個決策樹的結果，並且對於某些類別不平衡的數據集可能產生偏差。

```
clf = RandomForestClassifier()
    # Define the parameter grid
    parameters = {'n_estimators': [100, 200, 300],
        'min_samples_split': [2, 4, 6],
        'min_samples_leaf': [1, 2, 3],
        'max_depth': [10, 20, 30],
        # "criterion": ["gini"],
        'bootstrap': [True, False]
    }
```

4. GaussianNB:

- 優點：基於貝葉斯理論，具有計算效率高、易於實現和適應多類別問題的優點。
- 缺點：假設特徵之間相互獨立，可能無法處理特徵之間的關聯性。

```
clf = GaussianNB()  
    parameters = {  
        'var_smoothing': [1e-9, 1e-8, 1e-7, 1e-6, 1e-5]  
    }
```

5. GradientBoostingClassifier:

- 優點：通過組合多個弱學習器，具有良好的預測性能，並且能夠處理非線性特徵和大量數據。
- 缺點：模型訓練時間相對較長，對於噪聲數據和過擬合風險敏感。

```
clf = GradientBoostingClassifier()  
    parameters = {  
        'learning_rate': [0.01, 0.1],  
        'n_estimators': [50, 100],  
        'max_depth': [3, 5],  
        'min_samples_split': [2, 5],  
        'min_samples_leaf': [1, 2],  
        'max_features': ['sqrt', 'log2']  
    }
```

6. AdaBoostClassifier:

- 優點：通過迭代的方式提高模型的預測性能，對於不平衡數據集和噪聲數據具有較好的魯棒性。
- 缺點：對於極端噪聲和異常值數據比較敏感，可能導致過擬合。

```
clf = AdaBoostClassifier()  
    parameters = {  
        'n_estimators': [50, 100, 200],  
        'learning_rate': [0.1, 0.5, 1.0],  
    }
```



(三) 模型訓練和評估：

- 套件

```
# accuracy
from sklearn.metrics import accuracy_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score

# Model
from sklearn import tree
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn import tree
from sklearn.model_selection import GridSearchCV

# result
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
pd.set_option('display.max_columns', None)
```

- 建模方式：直接 import scikit-learn 中的 model。

- 調整方式：

對於機器學習模型的調參，Grid Search CV（交叉驗證網格搜索）是一種常用的方法，可以幫助我們找到最優的參數組合，從而提高模型的預測性能。在應用於新創公司成功與否的預測中，我們利用Grid Search CV來調整模型的參數，以獲得最佳的預測結果。以下是我們使用Grid Search CV進行參數調參的步驟：

1. 定義模型：選擇一個機器學習模型（如前述提到的AdaBoostClassifier）作為基礎模型。
2. 定義參數範圍：確定需要調整的參數範圍。
3. 網格搜索：使用Grid Search CV來搜尋參數空間中的最佳組合。網格搜索會遍歷所有參數的可能組合，對每個組合進行交叉驗證評估，並找到最優的參數組合。
4. 交叉驗證：在每個參數組合上進行交叉驗證，以評估模型的性能。這可以幫助我們估計模型在不同數據子集上的泛化能力，避免模型對特定數據的過擬合。
5. 模型評估：選擇具有最佳性能指標的參數組合作為最終模型。我們以準確率作為評估方式。

```
# Create a GridSearchCV model
grid_search = GridSearchCV(estimator = clf, param_grid = parameters,
                           cv = 5, n_jobs = -1, verbose = False, scoring="accuracy")
```

```
grid_search.fit(X_train, y_train)
best_model = grid_search.best_estimator_
best_params = grid_search.best_params_
```



（四）結果和討論：

利用新的資料集套入上述模型，以“Predict accuracy”評估我們所選模型在應用於未見數據時的準確性和有效性。

- 新資料集：
新資料集是獨立收集的，代表了一組不同的新創公司。這確保了我們的模型在超出其訓練數據範圍的真實情境下進行測試，從而提供了更真實的評估。
 - 資料預處理步驟：
為了驗證模型，我們遵循了與訓練階段相似的數據預處理過程。這包括處理缺失值、對特徵進行縮放和編碼類別變量，以確保與模型要求的一致性和兼容性。
1. Decision Tree：
在訓練集上的準確率為0.938，測試集上的準確率為0.934。最佳參數為{'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 3, 'min_samples_split': 5}。
 2. Gradient Boosting Classifier：
在訓練集上的準確率為0.938，測試集上的準確率為0.934。最佳參數為{'learning_rate': 0.01, 'max_depth': 3, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}。
 3. Ada Boost Classifier：
在訓練集上的準確率為0.938，測試集上的準確率為0.934。最佳參數為{'learning_rate': 0.1, 'n_estimators': 50}。
 4. GaussianNB：
在訓練集上的準確率為0.228，測試集上的準確率為0.238。最佳參數為{'var_smoothing': 1e-09}。
 5. RandomForestClassifier：
在訓練集上的準確率為0.938，測試集上的準確率為0.934。最佳參數為{'bootstrap': True, 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}。
 6. Logistic Regression：
在訓練集上的準確率為0.938，測試集上的準確率為0.934。最佳參數為{'C': 0.01, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear'}。

利用第二份資料集評估模型準確性：

1. Decision Tree:
 - Prediction accuracy: 0.7801418439716312
2. Gradient Boosting Classifier:
 - Prediction accuracy: 0.7257683215130024
3. Ada Boost Classifier:
 - Prediction accuracy: 0.7186761229314421
4. GaussianNB:
 - Prediction accuracy: 0.3546099290780142
5. Random Forest Classifier
 - Prediction accuracy: 0.7872340425531915
6. Logistic Regression:
 - Prediction accuracy: 0.6997635933806147

根據我們的研究結果，這些機器學習模型對於預測新創公司的成功與否具有相似的準確性。然而，GaussianNB模型的表現相對較差，可能是因為GaussianNB是一種假設特徵獨立性的朴素貝葉斯分類器，即它假設一個特徵的存在與其他特徵的存在無關。然而，在真實世界的情況下，新創公司成功預測數據集中的特徵往往存在相關性和依賴關係。這種獨立性的假設可能不成立，導致模型表現不佳。

此外，GaussianNB模型假設特徵的分布遵循高斯（正態）分布。如果數據集中的特徵不符合這種分布假設，它可能會影響模型的準確性。在新創公司成功預測中，特徵的分布可能偏離正態分布，進一步影響模型的性能。

此外，與其他算法（如決策樹或集成方法）相比，GaussianNB是一個相對簡單且不太靈活的模型。它可能難以捕捉數據中的複雜關係和模式，從而降低了預測準確性。

特徵重要性：

依據所繪出的特徵重要性圖，我們發現地區/是否為科技公司/募資對新創公司成功與否影響較大。

• 模型優點：

1. 自動化預測：機器學習模型能夠自動地從歷史數據中學習模式和關聯性，並將這些模式應用於新的數據，從而實現自動化的預測。這大大節省了時間和人力成本，並提高了預測的效率。

2. 能夠處理大量數據：隨著數據量的增加，傳統的人工方法往往變得困難且耗時。機器學習模型可以處理大量的數據，並從中學習並找出模式和趨勢，進而進行預測。這使得我們能夠更全面地考慮各種特徵和因素，從而提高預測的準確性。
 3. 靈活性和擴展性：機器學習模型具有良好的靈活性和擴展性，可以應用於各種不同的領域和問題。無論是預測新創公司的成功與否還是其他領域的問題，機器學習模型都可以通過調整參數和特徵選擇來適應不同的需求。
 4. 可解釋性：一些機器學習模型，如決策樹和邏輯回歸，具有較強的可解釋性，可以幫助我們理解模型是如何做出預測的。這對於理解預測結果的原因以及相應的行動計劃至關重要。
 5. 無偏見：機器學習模型在預測中不受主觀偏見的影響。基於數據進行學習和預測，避免了人為偏見的介入。這使得預測更加客觀和可靠。
- **模型局限性**：模型的預測結果仍然受到數據質量和特徵選擇的影響。此外，過去的數據不一定能完全預測未來的結果，因為市場和經濟環境的變化可能會對新創公司的成功產生重大影響。
 - **應用前景**：我們的研究結果表明，利用機器學習模型對新創公司的成功與否進行預測是可行的。透過適當的特徵選擇和模型調參，我們能夠提高預測的準確性和可靠性，從而為投資者、創業者和風險投資基金等利益相關者提供更明智的決策支持。



（五）遇到的困難和解決方法：

1. 數據不完整或缺失：新創公司的數據可能存在缺失值或不完整的情況，這會對模型的性能產生不利影響。我們解決這個問題的方法是進行數據清理和預處理，例如填補缺失值、處理異常值。
2. 不平衡的數據集：在新創公司成功與否的數據集中，成功公司和失敗公司的比例可能不平衡。這會導致模型在預測時對多數類別（如成功公司）的預測性能更好，而對少數類別（如失敗公司）的預測性能較差。解決這個問題的方法包括適當的採樣方法（如過採樣或欠採樣）或使用相應的評估指標（如F1分數）來評估模型的性能。
3. 不確定性：任何預測模型都存在不確定性。即使模型在訓練集上表現良好，也無法保證在新數據上能夠完全準確預測。未來解決這個問題的方法是使用機器學習模型的不確定性估計方法，如置信區間、概率預測等，以提供預測的可信度量。

4. 維護和更新：預測模型需要定期維護和更新，以應對數據和環境的變化。隨著新數據的累積和模型性能的評估，未來可能需要重新訓練模型、調整參數或重新選擇特徵。



（六）心得感想：

通過這次在預測新創公司成功與否的應用中的實踐，我深切體會到機器學習的強大和應用的價值。這項任務不僅涉及到數據處理和特徵選擇等技術層面的挑戰，更需要對新創公司的相關特徵和成功因素有深入的理解。

在解決困難的過程中，我們學到了許多寶貴的經驗和技巧。例如，數據清理和預處理是確保模型性能的重要步驟，特別是在面對數據不完整和不平衡的情況時。

同時，模型的選擇和調參也是影響預測準確性的重要因素。不同的模型有不同的優勢和限制，需要根據具體情況選擇適合的模型。調優參數則需要通過交叉驗證等技術進行實驗和優化，以提高模型的性能。

除了技術層面，我們也更加深入地理解了預測模型的局限性和不確定性。即使模型在訓練集上表現良好，也無法保證在新數據上的預測能力。因此，對模型的結果要有一定的警惕性，並時刻意識到預測的不確定性。

總而言之，通過這份報告，我們發現機器學習模型在預測新創公司成功與否方面具有重要的應用價值。它能夠處理大量數據，具有靈活性和擴展性，並提供了客觀和可靠的預測結果。然而，我們也要注意模型的不確定性和持續維護的重要性。這個項目讓我們深入了解了機器學習在預測問題中的應用，並通過解決困難和挑戰來提高模型的性能。我們相信這項技術在未來將會在投資、創業和風險評估等領域發揮重要作用。同時，我們也深刻體會到持續學習和改進模型的重要性，因為數據和環境的變化可能需要我們不斷調整和優化模型，以保持預測的準確性和可靠性。



（七）未來展望：

雖然我們的研究利用機器學習技術預測新創公司的成功具有價值的洞察，但在幾個方面仍然可以進一步改進。旨在提高模型的準確性和應用性：

1. 特徵工程：探索更高級的特徵工程技術，以捕捉有關新創公司更多相關的信息。這可以包括提取特定領域的特徵、創建交互項，或者整合外部數據來豐富特徵集。
2. 模型集成：考慮通過模型集成技術（如堆疊、混合或提升）結合多個模型的預測。這有可能提高預測準確性，減少過擬合的風險。


3. 處理不平衡數據：制定有效的策略來處理數據集中的類別不平衡問題。可以使用過採樣少數類別、欠採樣多數類別或使用混合採樣方法等技術，以解決這個問題並提高模型的性能。
4. 結合時間序列分析：如果有相應的數據，考慮結合時間序列分析技術來捕捉新創公司成功的時間模式和趨勢。這可以提供更深入的洞察，使預測更準確。
5. 模型可解釋性：探索提高模型可解釋性的方法，同時保持預測性能。例如，特徵重要性排序、部分依賴圖或代理模型等技術可以提供有關影響預測的因素的洞察，有助於決策過程。
6. 持續模型監控：建立一個持續監控和更新預測模型的系統。隨著新的數據變得可用和新的創業環境的變化，模型的性能可能會發生變化。定期重新訓練和評估模型將確保其在時間上的相關性和準確性。
7. 風險預測：除了成功與否的預測，將注意力轉移到預測新創公司的風險和失敗的可能性。這將幫助投資者和利害相關者更全面地評估創業機會和風險，並制定相應的策略。
8. 人工智能應用：將其他人工智能技術應用於預測新創公司成功，例如自然語言處理（NLP）分析創業者的文字資訊、圖像處理分析新創公司的品牌形象等。這將擴大預測模型的視野和能力。
9. 影響因素研究：進一步研究影響新創公司成功與否的因素，包括市場環境、資金結構、行業競爭等。這將有助於全面理解創業環境和成功因素的相互關聯性。

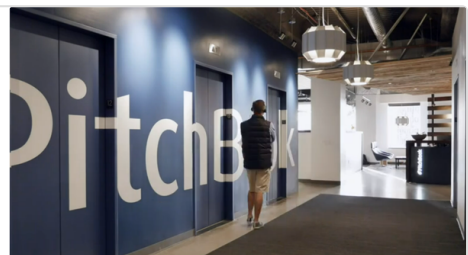


（八）References：

PitchBook用AI預測新創公司能否成功活下去

AI真的很好用，除了可以幫人寫論文外，還能演算出一家新創公司能不能活下去，PitchBook就出了這樣一套AI。


 <https://newsforchinese.com/v3/2023/03/pitchbook用ai預測新創公司能否成功活下去/>



sklearn.model_selection.GridSearchCV

Examples using sklearn.model_selection.GridSearchCV:

Release Highlights for scikit-learn 0.24 Release Highlights for scikit-learn 0.24 Feature agglomeration vs. univariate selection


 https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html



sklearn.naive_bayes.GaussianNB



Examples using sklearn.naive_bayes.GaussianNB: Comparison of Calibration of Classifiers Comparison of Calibration of Classifiers Probability Calibration curves Probability Calibration

 https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html