

計畫名稱：Affordable Credit Classifier System (ACCS)

第一章 創新構想

創造一個可供大眾查詢個人信用狀況之簡易平台，用於借貸、網拍、商業合作來評估對方的信用，來決定合作可行性，評分方式用過去大數據建構的資料庫作為訓練數據，對有需要的人進行適當的信用分類，降低維護成本。

第二章 金融應用領域說明

1. 產品或服務現有缺失

舊機關維護成本高，從聯合徵信中心可以看到，一般查詢成本高，一年享有一次免費查詢，第二次(含加查)以後開始將酌收成本費用，線上查閱80元/次，紙本報告第二次以後100/次。

2. 產品或服務創新方案

提供一個可供大眾查詢個人信用狀況之簡易平台，並用過去大數據建構的資料庫作為訓練數據，對未來新的人進行適當的信用分類，降低維護成本。

信用評分將會有多項準則，繳款行為類信用資料，負債類信用資料，其他類信用資料為三大準則：

● 繳款行為類信用資料

個人過去在信用卡、授信借貸(信貸、學貸、各種貸款)以及票據(匯票、本票及支票)的還款行為表現，目的在於瞭解個人過去有無不良繳款紀錄及其授信貸款或信用卡的還款情形，主要包括其延遲還款的嚴重程度、發生頻率及發生延遲繳款的時間點等資料。

● 負債類信用資料

1. 個人信用的擴張程度，主要包括負債總額，例如：信用卡額度使用率=應繳金額加上未到期金額÷信用卡額度，或是授信借款往來金融機構家數
2. 負債型態，例如：信用卡有無預借現金、有無使用循環信用、授信有無擔保品
3. 負債變動幅度，例如：授信餘額連續減少月份數。
4. 互評制度：使用戶之間可以依據自身合作經驗，進行互相評分，當作參考依據。

第三章 機械學習運算法應用說明

1. 資料收集

我們在 Kaggle 網站上找到了一個名為 Credit score classification 的資料集。這個資料集收集了一家全球金融公司多年來收集的基本銀行詳細資料和大量信用相關資訊。該公司管理層希望建立一個智能系統，將人們分類為信用分數

段，以減少人工操作。任務是給定一個人的信用相關資訊，建立一個能夠將信用分數分類的機器學習模型。

2. 資料特徵

該資料集的目標是預測 Credit_Score，這是一個分類變量，可以是 Good、Standard或Poor。此外還有包含22種有意義的特徵：

- ID : ID
- Age : 年齡
- Occupation : 職業
- Annual_Income : 年收入
- Monthly_Inhand_Salary : 固定月收入
- Num_Bank_Accounts : 銀行帳號的數量
- Num_Credit_Card : 信用卡持有的數量
- Interest_Rate : 信用卡的利率
- Num_of_Loan : 跟銀行有級筆貸款
- Type_of_Loan : 貸款的類型
- Num_Credit_Inquiries : 信用卡查詢次數
- Delay_from_due_date : 平均拖欠款項的天數
- Credit_Mix : 信貸類型
- Outstanding_Debt : 在外流通債金額
- Credit_Utilization_Ratio : 信用卡使用率
- Credit_History_Age : 過去使用信用的年數
- Payment_of_Min_Amount : 只有最低限度繳回
- Total_EMI_per_month : Equated Monthly Installment
- Amount_invested_monthly : 每個月的投資金額
- Payment_Behaviour : 付款習慣
- Monthly_Balance : 每月帳戶餘額
- Num_of_Delayed_Payment : 平均一人拖欠貸款數量

3. 資料前處理

在資料前處理步驟中，我們首先移除了一些無用的特徵: 'ID' 'Name' 'SSN'，接下來，我們將數值都轉換為 float 類型，並將字串類的設為 "object" 類型。

接下來，我們處理了缺失值，KNNImputer 是一種用於處理缺失值的方法，其中 n_neighbors 參數決定了 KNN 演算法中的 K 值，當 n_neighbors=1 時，KNNImputer 會找到缺失值最近的鄰居，並使用該鄰居的值來填補缺失值，這樣就可以在不丟失太多資訊的情況下，將缺失值填補完整。

然後，我們對 "object" 類型的特徵進行 One-Hot Encoding，它是一種將類別變量轉換為數值變量的方法，它會將每個類別變量拆分成一個二元變量，其中只有一個變量的值為 1，其餘變量的值都為 0。例如，如果有一個類別變量包

含三個類別 A、B 和 C，則在 One-Hot Encoding 後，該變量將被拆分成三個二元變量，分別表示 A、B 和 C。One-Hot Encoding 對機器學習有很大的優點，最大的好處是它可以讓我們使用許多機器學習模型，因為這些模型通常只能處理數值類型的變量。

在處理不平衡資料的方面我們使用了 SMOTE (Synthetic Minority Oversampling TEchnique)，它可以通過生成新的合成樣本來平衡分類資料集中的類別。SMOTE 演算法首先會找到資料集中少數類別的樣本，並在其周圍生成新的合成樣本。生成的合成樣本是在少數類別樣本與其他鄰近樣本之間隨機插值得到的。這樣就可以增加少數類別的樣本數，並使分類資料集平衡。SMOTE 演算法常用於資料挖掘和機器學習領域，可以有效改善在不平衡資料集上訓練的模型性能。

最後，我們將訓練資料集分割為 Train (80%) 和 Test (20%)。Train 集將用於訓練模型，Test 集將用於評估模型的性能。這樣可以幫助我們檢驗模型是否具有良好的泛化能力，即能夠適用於未知資料。

4. 機器學習方法

我們實驗了以下方法：

- **LogisticRegression**

Logistic Regression 是一種用於做二元分類的監督式學習演算法。它通過在特徵空間中找到一個超平面，將樣本分類到兩個類別中。Logistic Regression 常用於預測類別變量，如預測一個人是否有某種疾病。

- **KNN**

KNN (K Nearest Neighbors) 是一種基於定義相似性的監督式學習演算法。它會找出最近的 K 個鄰居，並使用多數決的原則將新的樣本分類到最多數的類別中。KNN 常用於分類任務，並且它的性能會受到 K 值的影響。

- **Decision Tree**

Decision Tree 是一種決策樹演算法，可以在特徵空間中找到一棵用於分類的樹。它會根據資料集中的特徵，建立決策樹模型。在每個節點，決策樹會根據最佳分割特徵將資料分類到子節點中。Decision Tree 常用於分類和回歸任務。

- **Random Forest**

Random Forest 是一種集成學習演算法，它會建立多棵 Decision Tree，並使用多數決的原則將新的樣本分類到最多數的類別中。Random Forest 可以處理分類和回歸任務，並且它的性能通常比單棵 Decision Tree 要好。

- **XGBClassifier**

XGBClassifier 是一種基於梯度提升演算法的分類器。它可以快速訓練複雜的模型，並且可以處理大量的資料和特徵。XGBClassifier 可以幫助我們建立高效率的分類模型，並且在許多應用中表現出良好的性能。

- **AdaBoostClassifier**

AdaBoostClassifier 是一種基於自適應提升演算法的分類器。它會建立多個弱分類器，並將它們的預測結果結合起來。AdaBoostClassifier 會不斷修正分類器的權重，以便讓分類器更加精確。AdaBoostClassifier 可以處理分類和回歸任務，並且在許多應用中表現出良好的性能。

- **Support Vector Machine**

Support Vector Machine (SVM) 是一種監督式學習演算法，可以用於分類和回歸任務。SVM 會在特徵空間中找到一個超平面，使得兩個類別的資料有最大的間隔。SVM 常用於資料的高維度空間中，並且它的性能通常比其他算法要好。

5. Feature Selection

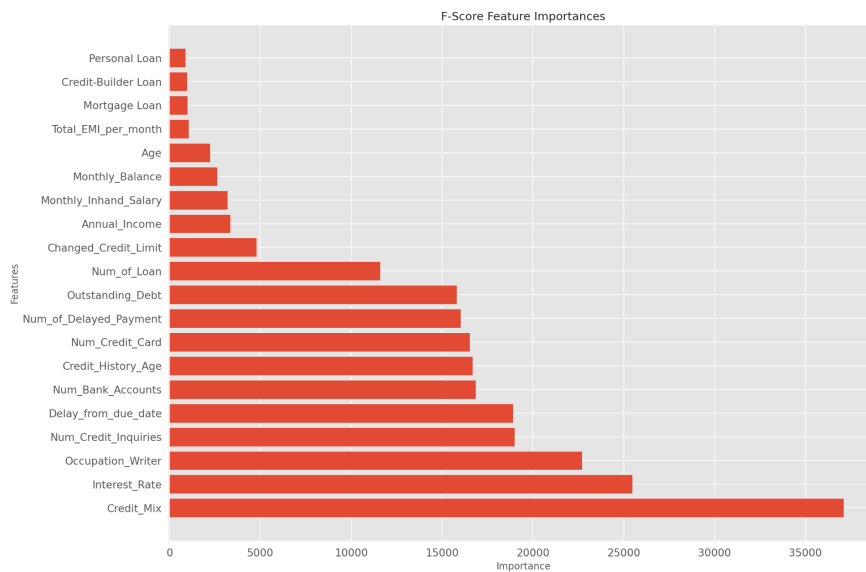
F-score 是一種 Filter 方法，用於評估每個特徵與目標變量之間的相關性。它會計算每個特徵的 F-score，並根據 F-score 排序。F-score 是由精確率和召回率的調和平均數計算而來，公式如下：
$$F\text{-score} = 2 * (\text{精確率} * \text{召回率}) / (\text{精確率} + \text{召回率})$$
使用 f-score 的 Feature Selection 方法，我們可以選擇 F-score 高的特徵。選擇 F-score 高的特徵可以提高模型的性能，因為這些特徵與目標變量有較強的相關性。使用 f-score 的 Feature Selection 方法的一個缺點是，它只能評估每個特徵與目標變量之間的相關性，無法評估特徵之間的相互作用。因此，f-score 可能會忽略一些重要的特徵。

6. 實驗結果

- 全資料當作特徵

Classifier	Train Acc	Test Acc
LogisticRegression	72.06%	71.52%
KNeighborsClassifier	89.55%	82.44%
DecisionTreeClassifier	100.00%	74.26%
RandomForestClassifier	100.00%	83.39%
XGBClassifier	87.55%	83.67%
AdaBoostClassifier	73.37%	73.33%
Support Vector Machine	74.32%	73.62%

- Feature Selection



- 以Feature Selection top 10 當作特徵

Classifier	Train Acc	Test Acc
LogisticRegression	69.73%	69.33%
KNeighborsClassifier	88.88%	83.92%
DecisionTreeClassifier	99.83%	83.02%
RandomForestClassifier	99.83%	87.56%
XGBClassifier	81.94%	79.78%
AdaBoostClassifier	71.69%	71.42%
Support Vector Machine	72.81%	72.53%

- 以Feature Selection top 20 當作特徵

Classifier	Train Acc	Test Acc
LogisticRegression	70.28%	70.41%
KNeighborsClassifier	89.52%	84.90%
DecisionTreeClassifier	100.00%	82.38%
RandomForestClassifier	99.99%	88.19%
XGBClassifier	84.74%	82.02%
AdaBoostClassifier	72.17%	72.28%
Support Vector Machine	72.62%	72.74%

第四章 創新營運模式

簡述: 創造一個會員報名制度, 透過資料集及即演算法, 去篩選社會地位高的人才, 以繳交年費可自由參加, 主要目的是拓展人脈、增進自己, 會依據信用評分(負債率、帳戶餘額等等)、年收入、投票制度去分成鑽石會員(高等)、白銀會員(初等)。在大數據時代, 數據是非常有價值的物品, 我們想藉此利用客戶的數據將營運方式最大化利益。

- **鑽石會員優勢:** 在確保大家都是年收高, 信用評分高的情況下, 定期舉辦活動讓鑽石會員可以拓展人脈 (Networking), 例如:研討會、打高爾夫球、酒店聚餐、出國旅遊、露營。且定期舉辦課程, 讓會員可以上股票分析課、品酒課、藝術探討。(每個活動均有報名費)
- **白銀會員優勢:** 主要以上專業課程為主, 目的為了增進自己價值, 像是專業理財投資課、管理顧問課、專業證照課。(每個課程均有報名費)

第五章 結論與預期效益



預期效益: 1.ACCS:平台給予大眾簡單, 快速, 便利的來查詢信用評分, 在創造便利的同時, 第二次查詢向民眾收取一百元。

2.廣告收益:查詢過程中添加數個廣告, 不想看廣告的人, 我們會予以收費來去除廣告。

3.會員年費:鑽石與白銀會員繳交的費用, 和Networking 活動報名費&專業課程費, 對會員舉辦社交聚會與課程來讓他們聯繫感情, 同時我們也能從中獲利。

結論:創造一個可供大眾以低成本查詢信用狀況之簡易平台, 用於借貸、網拍、商業合作來評估對方的信用(只顯示部分資料), 來決定合作可行性, 平時的業務由基本的搜尋計費以及廣告收益, 除此之外, 我們希望利用收集到的資料去建立一個會員制度, 希望能藉此創造一個平台讓信用好、年收高、想學習增進自己的人加入會員, 讓會員可以與其他篩選過的會員進行交際、活動、研討會以及專業課程等等, 形成一個互惠互利的環境。

第六章 參考文獻

- **Data :** Credit score classification | Kaggle
- **Data Processing :** Credit Score EDA  + Prediction (Multi-Class)  | Kaggle
- **Model :** Examples — scikit-learn 1.2.0 documentation
- **Feature Selection :** SK Part 2: Feature Selection and Ranking
- **Jupyter Notebook :** Project Jupyter | Home
- **Google Colab :** Google Colab