**40871223H 徐存昇**

**Data Mining Final Project**

報告目的：

比較RFC、XGBOOST、SVM對該資料的適合程度評估

動機：

由於model有許多種，我希望挑選出大宗主流的模型進行比較，因此從：

梯度提升機：XGBOOST

支持向量機：SVM

決策樹集成學習：RFR

來進行比較

1. 預處理1  
   將全部資料的600份隨機亂數取成Standard Answer。剩下的資料當作測試集，並且加以訓練，之後的預測結果都會跟Standard Answer比較相似度。當作預測成績。
2. 預處理2  
   判斷是否有Null值，發現Year跟Publisher都有缺失。  
   因此Year取平均2006年，Publisher補None。  
   一張含有 文字, 螢幕擷取畫面, 字型 的圖片

   自動產生的描述
3. 預處理3  
   將文字型態轉化為數字，方便之後Model進行訓練。  
   一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

   自動產生的描述
4. 機器處理：隨機森林

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

調整參數程式：

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

RFR調參前

一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述

RFR調參後

一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述



可明顯發現調整參數後，平方交叉的成績有上升。

但是我們將檔案輸出後，卻發現與原始檔案相思度卻降低了。

1. 機器處理：XGBOOST

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

XGBOOST

一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述

1. 機器處理：SVM

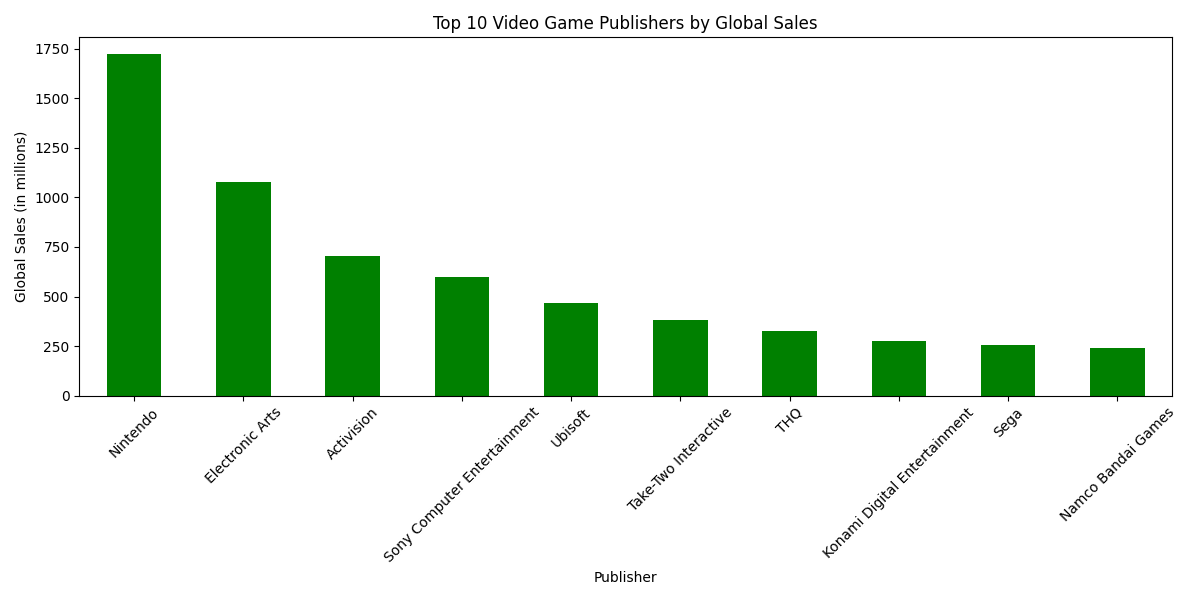
一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

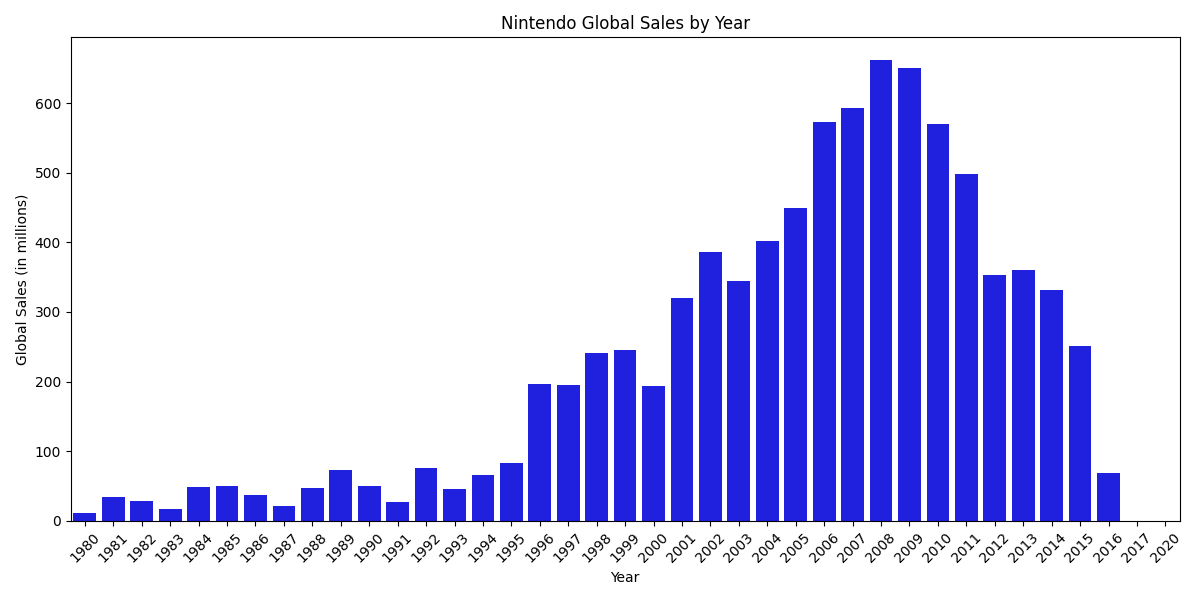
SVM



1. 前十大遊戲公司排名  
   一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

   自動產生的描述  
   
2. Nitendo的營收狀況

一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

自動產生的描述

1. 結論

該數據擁有約16600個數據，算蠻大的數據量，在預處理的部分有缺失值[Year]和[Publisher]，年份取平均值填補為2006，發行商則用Unknown來進行，最後預測前利用LabelEncoder轉換為數字進行預測。

比較了支持向量機(SVM)、梯度提升機(XGBOOST)、決策樹回歸(RSR)三種大類型對該數據的比較。理論上，SVM適合高特徵、量不大的數據，顯然與我們資料相悖，因此支持度也是最低的；而XGBOOST是適合大型數據，因此有不錯的支持度支撐；RFR可以承受較高的過擬合風險，因為是多棵決策樹，也適合應用於較大的數據量。最終結果是RFR擁有較好的相似度。

調整參數的過程中，會發現較高的cross\_val反而在相似度比對其實沒有比較高，有可能達到了過擬合的狀態。