Personal Final Project

姓名：何子安

學號：E44065020

系級：心理 110

任務簡介

針對自行選定的資料集題目 – 電信公司顧客留存預測。以顧客的相關資料或行為來預測最終顧客是否會離開或繼續留在該公司，是一個二元分類任務。在這之中也會對整個資料集進行盡可能全面的資料分析與資料探勘，並於報告中呈現我們的發現。

資料概述

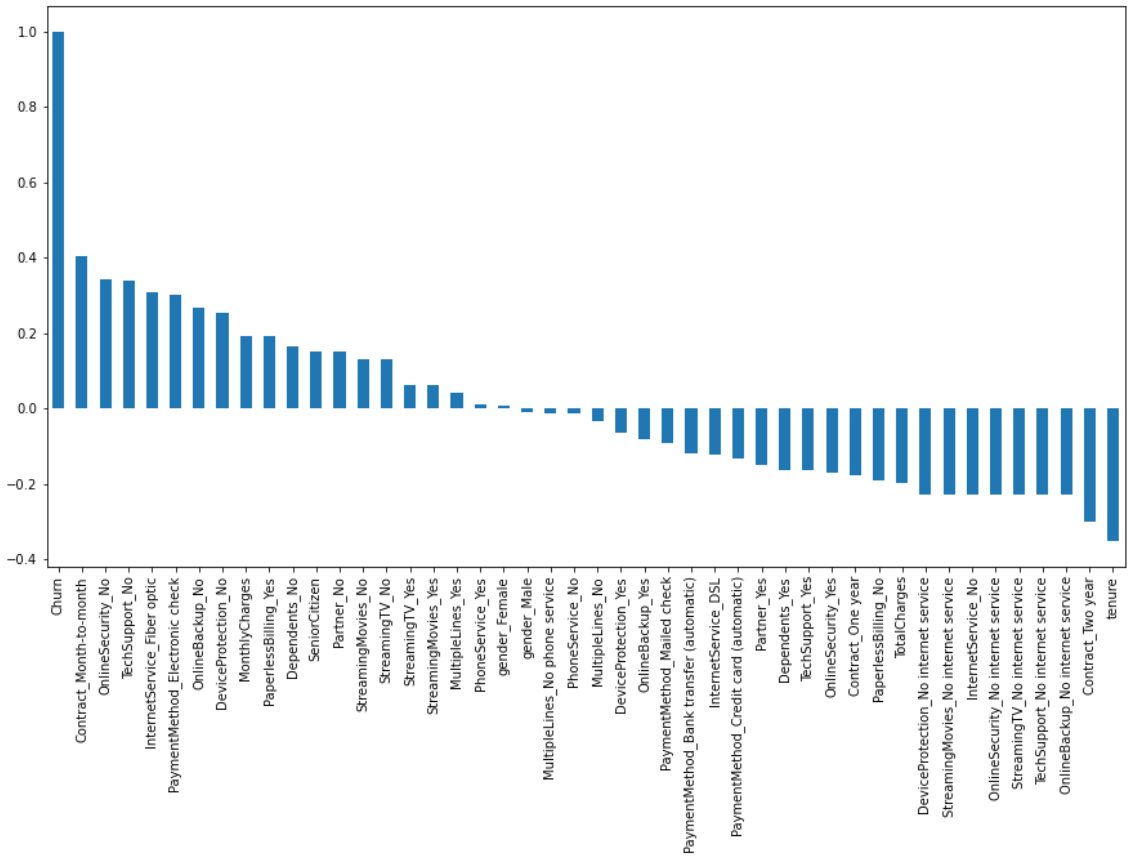
本次分類任務是於 GitHub 上發現的。在本次任務中只有一份資料集，在該資料集中的每一行所代表的是每一位顧客，而每一列所代表的是該顧客的屬性、特徵或有過的行為。以下簡述該資料集中有包含的資訊（各欄位）：

- 【churn】在上一個月中確實離開公司的顧客。
- 【phone, multiple lines, internet】各顧客是否有向公司簽下的各種服務。
- 【contract, payment method, paperless billing, monthly charges, total charges】關於顧客個人帳戶中的一些資訊。
- 【gender, age range, partners, dependents】關於顧客私人或敏感的資訊。

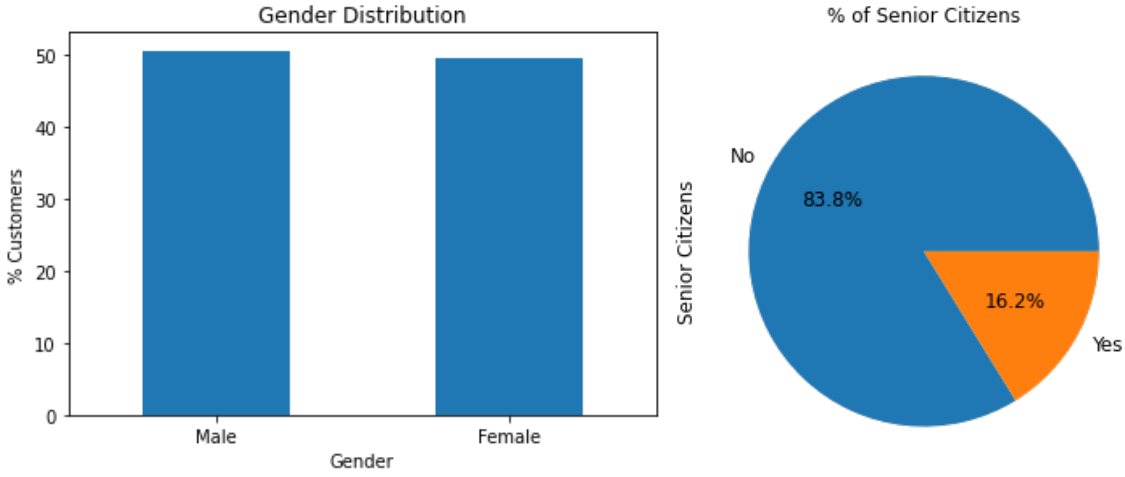
資料分析

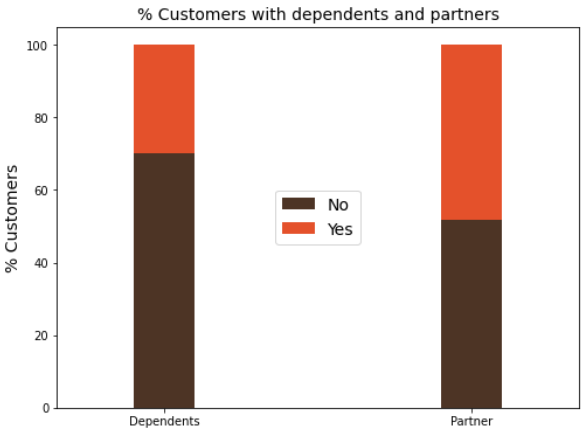
整個資料集含有 21 個欄位，並有 7043 筆顧客的資訊，且沒有任何的缺失資料。以下為對整個資料集進行各種資料分析與資料挖掘，希望可以為這份資料集找出更出代表性的表示。

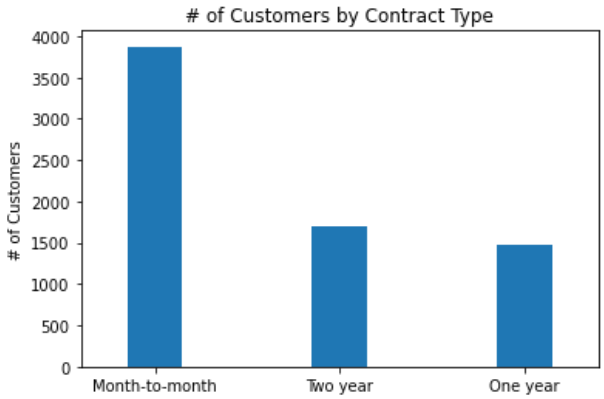
- 【churn】與其他特徵項之間的相關性：

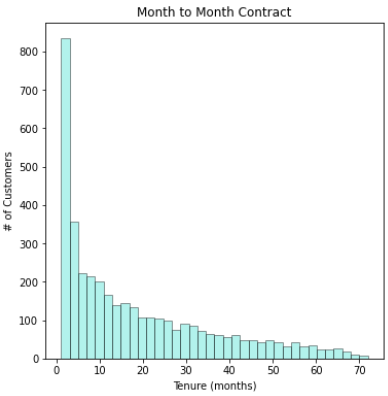


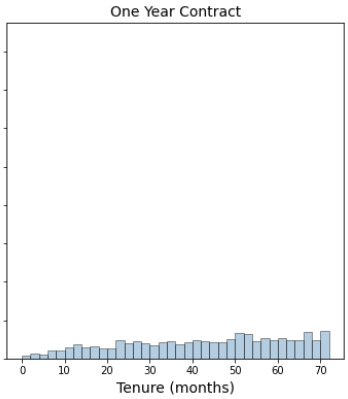
- 各種對資料集的資料探勘

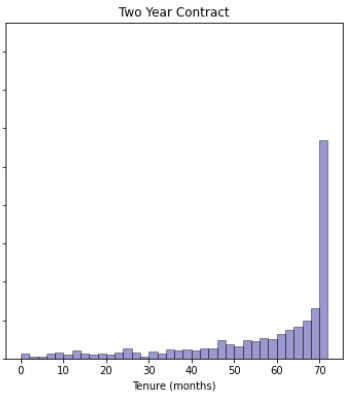


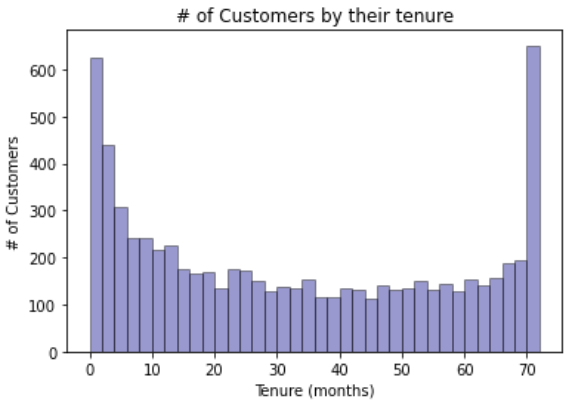

% Customers with dependents and partners
% Customers
No
Yes
Dependents
Partner

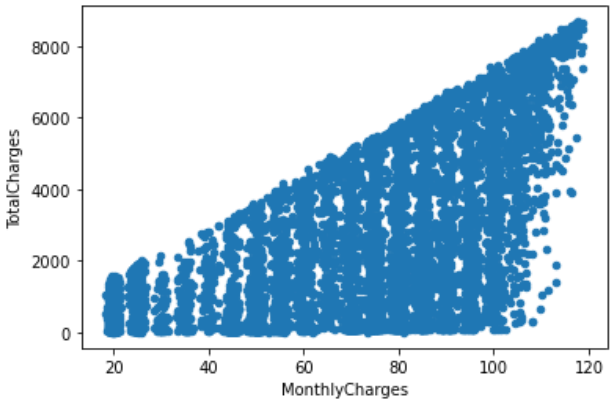

# of Customers by Contract Type
# of Customers
Month-to-month
Two year
One year

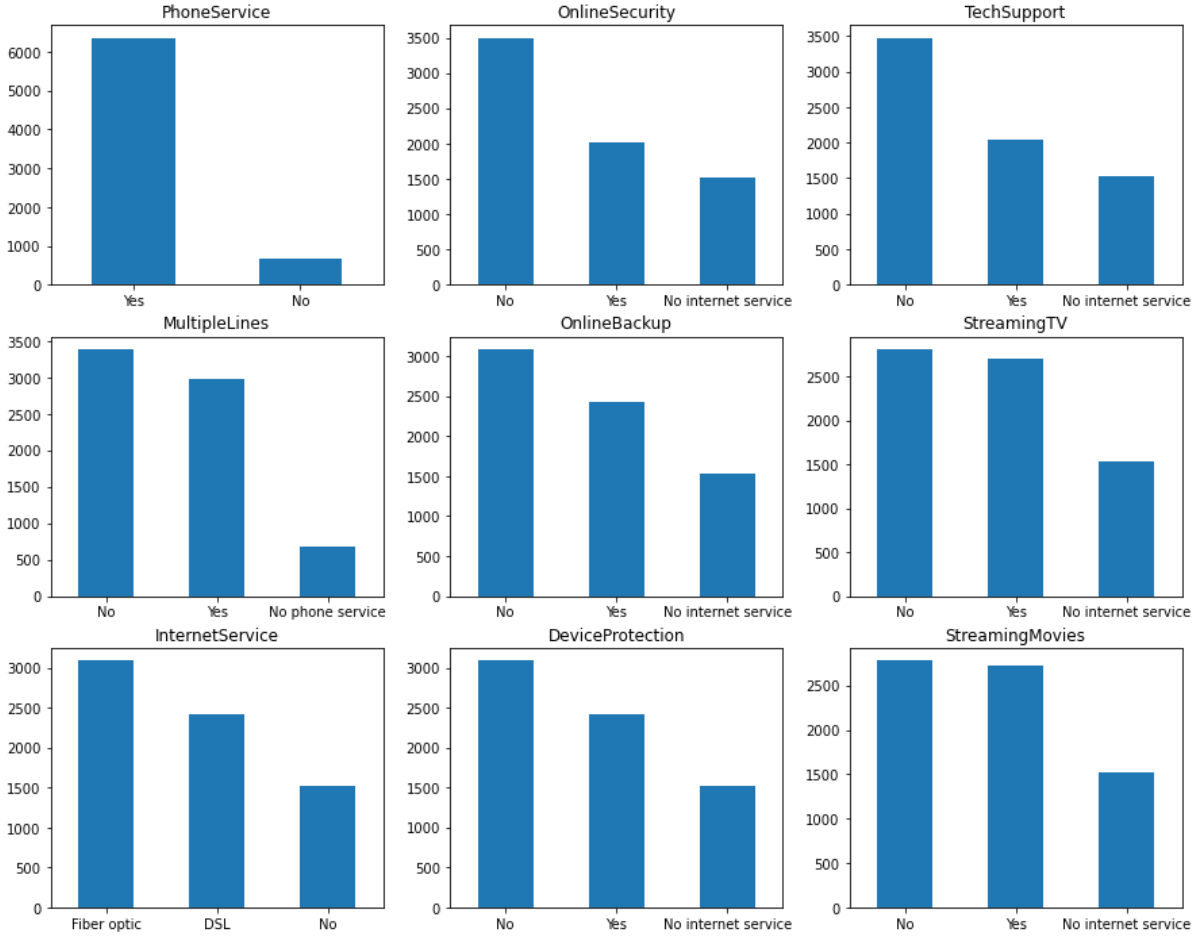

Month to Month Contract
# of Customers
Tenure (months)

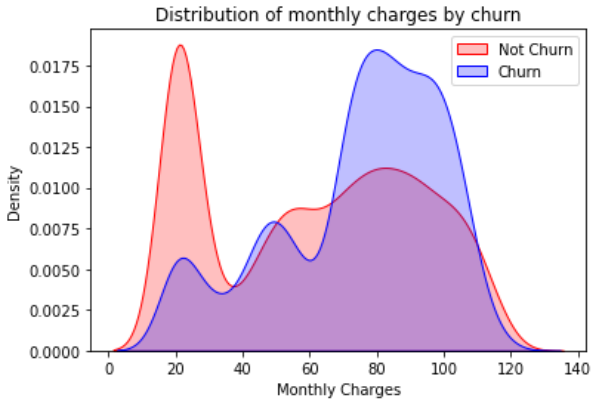

One Year Contract
Tenure (months)

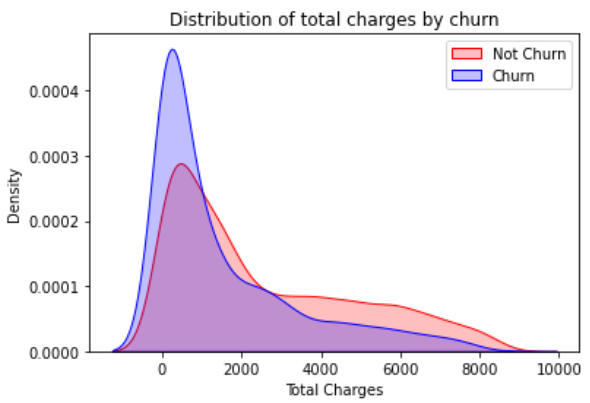

Two Year Contract
Tenure (months)

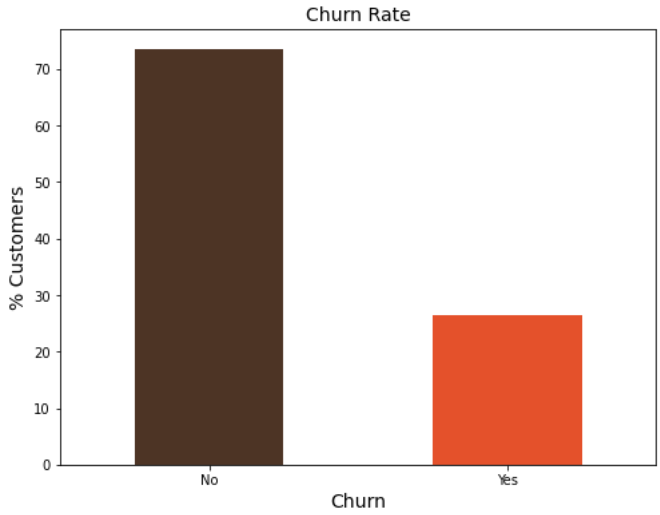

# of Customers by their tenure
# of Customers
Tenure (months)


TotalCharges
MonthlyCharges


PhoneService
Yes
No
OnlineSecurity
No
Yes
No internet service
TechSupport
No
Yes
No internet service
MultipleLines
No
Yes
No phone service
OnlineBackup
No
Yes
No internet service
StreamingTV
No
Yes
No internet service
InternetService
Fiber optic
DSL
No
DeviceProtection
No
Yes
No internet service
StreamingMovies
No
Yes
No internet service


Distribution of monthly charges by churn
Not Churn
Churn
Density
Monthly Charges


Distribution of total charges by churn
Not Churn
Churn
Density
Total Charges



資料前處理

在將資料特徵輸入各個模型之前需要確認該資料是否有符合我們的需要，如果不符合，就需要將對應的資料轉換為適合模型讀取的格式，而這一步驟稱之為「資料前處理」。首先，資料集中的【TotalCharges】所顯示的是文字型的資料，必須將其轉換為數值型資料。而對於資料集中用來代表顧客去留的欄位【Churn】的資料內容是 Yes/No，這並不適合輸入模型去讓模型進行學習，所以也將其轉換為對應的數值資料（將 Yes 取代為 1，No 取代為 0）。而在資料集中還存在其他的非二元類的數值型資料，則對該些資料欄位都用虛擬變數來表示。接著選擇使用了 MinMaxScaler 來對整個資料集進行調整，目的是將資料集中所有的數值都落在 0 和 1 之間。最後則是將資料集以 7：3 的比例切分為訓練資料與測試資料。

模型介紹

在模型的選擇上使用了各種較經典且熟悉的分類預測模型，並在此基礎上盡可能改良或找出最佳的超參數。且在模型解釋能力與模型複雜度之間取捨選擇使用了 Logistic Regression 、 Random Forest Classifier 、 Support Vector Machine 、 AdaBoost Classifier 來作為預測模型。

結果與討論

在本次分類任務中皆採用了 Accuracy 準確度作為評估模型效果之工具。 Logistic Regression 在這份資料集上進行預測的分數為 0.81；Random Forest Classifier 得到 0.80；Support Vector Machine 的準確度為 0.79；而 AdaBoost Classifier 則得到 0.81 的準確度。值得提的地方是竟然 AdaBoost Classifier 的表現也沒有特別好，甚至與 Logistic Regression 相差無幾。通過查看模型所預測結果的混淆矩陣，可以發現其實因為資料是呈現不平均的狀態，導致模型沒有很好的「學習」到資料的表示或分佈。

未來工作

考慮到資料的目標項分佈不均，未來可以加入增多資料量使得其分佈平衡的特徵工程（up-sampling 上取樣）來加強。另外，對於分類問題的預測模型也可以使用類神經網路來預測，在未來展望中應嘗試使用神經網路來訓練。最後，各種模型都有許多的超參數會影響模型的表現，可以考慮使用網格搜索（Grid Search）來為超參數找出最佳的值。

參考資料

- https://github.com/IBM/telco-customer-churn-on-icp4d
- sklearn.ensemble.AdaBoostClassifier
- sklearn.ensemble.RandomForestClassifier