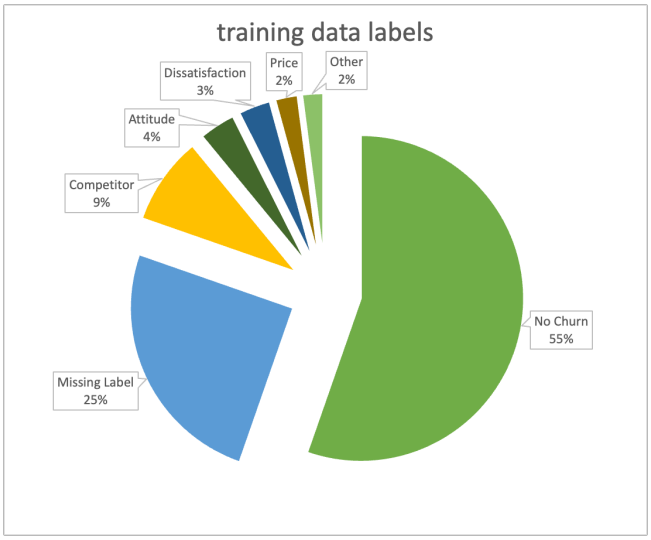
Final Project Report

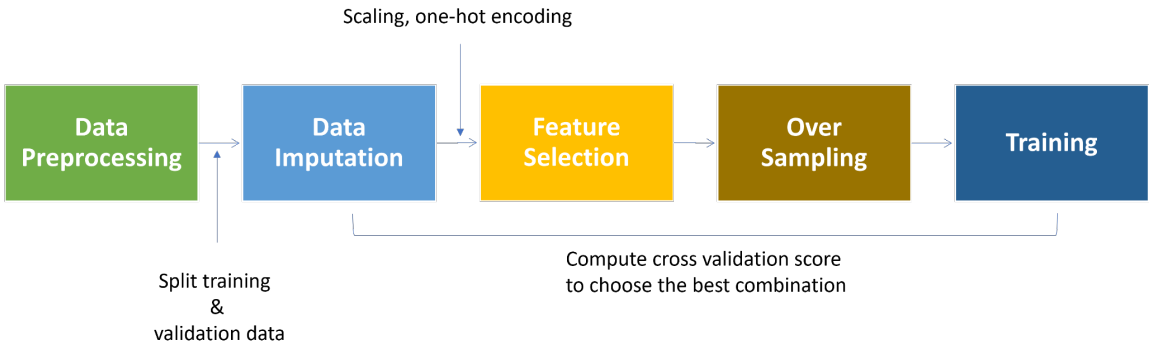
Team Name: Have you done your homework
Members: R10725023 吳琦艾、R10725045 蔡祐琳、R10922078 賴侃軒

1. Introduction

本次研究目標為預測電信服務的用戶會不會流失，以及可能的流失原因，期望能透過收集的用戶資料提前預知客戶的續約決策。資料集包含 6 個檔案，扣除 key value 及用於計算的欄位，feature 總計共 43 個，且缺失值約 25%；另外 label 存在不平衡的情況，會續約的客戶(No Churn) 佔比高達 73.7%，因此 data preprocessing 會以不同方法嘗試改善這些問題。本次研究以 F1-score 作為模型表現評估指標。



2. Experiment Procedure



本研究流程包含 Data Preprocessing、Data Imputation、Feature Selection、Over Sampling、Training。首先，Data Preprocessing 會將原始資料進行處理，使用邏輯補值的方式先初步進行處理，之後將資料分成訓練資料集及驗證資料集，之後進行缺失值的 Data Imputation。先使用訓練資料集去擬合 Imputation Method 後，再 transform 於驗證資料集。完成後進行 Scaling, one-hot encoding。另外，由於 encoding 過後 feature 的欄位超過1,000個，因此加入 Feature Selection，選出較具代表性的特徵進行模型訓練，避免雜訊過多及訓練時間過

久。在參考前述資料分析時發現 label 存在不平衡的情況，因此加入 Over Sampling 平衡訓練資料集。資料經過前述處理後，丟入訓練模型，並計算各種 Data Imputation、Feature Selection、Over Sampling、Training 的組合後，選出 F1 的 cross validation score 最好的上傳。

3. Approach Introduction and Detail

3.1. Data Preprocessing

首先，針對原始的 csv 進行邏輯補值。在 demographics.csv 的部分，「性別 (Gender)」的欄位將男性與女性用戶分別替換成 0 與 1 的類別；「年齡 (Age)」、「是否低於 30 歲 (Under 30)」、「是否高於 65 歲 (Senior Citizen)」的欄位，使用年齡判斷該用戶是否低於 30歲，及是否高於 65 歲，另外也透過比對「是否低於 30 歲」及「是否高於 65 歲」兩個欄位，使用有完整資料的用戶的平均年齡推測缺失的用戶年齡；「是否有扶養親屬 (Dependents)」、「扶養親屬的人數 (Number of Dependents)」兩個欄位也是使用與上述類似的方式進行交叉比對來補值。

在 services.csv 的部分，「是否曾推薦朋友使用本公司服務 (Referred a Friend)」及「推薦的朋友人數 (Number of Referrals)」的欄位使用與上述類似的方式進行交叉比對補值。透過「使用的網路服務類型 (Internet Type)」及「平均每月下載 GB 數 (Avg Monthly GB Download)」兩個欄位，可推測出此用戶「是否使用本公司的網路服務 (Internet Service)」；透過「續約月數 (Tenure in Months)」、「月租費 (Monthly Charge)」、「總收費 (Total Charges)」可交叉推算出各自的月數及費用；「是否使用數據吃到飽服務 (Unlimited Data)」與「總額外收費 (Total Extra Data Charges)」也可以互相驗證；其餘的欄位若為數值欄位，則使用平均及中位數兩種方式補值，若為類別欄位，則使用相同條件的用戶資料取眾數進行補值，如「使用的網路服務類型 (Internet Type)」欄位會使用「是否使用本公司的網路服務(Internet Service)」欄位相同的用戶來取眾數進行補值。

在location.csv的部分，使用「緯度 (Latitude)」及「經度 (Longitude)」推測出「用戶所在的城市 (City)」及「郵遞區號 (Zip Code)」，另外也使用「郵遞區號 (Zip Code)」配合population.csv找出用戶所在城市的「人口數 (Population)」。status.csv中的「用戶流失類別 (Churn Category)」欄位為本次的分析標的，將「用戶沒有流失 (No Churn)」、「其他電信業者因素 (Competitor)」、「用戶滿意度低 (Dissatisfaction)」、「服務態度 (Attitude)」、「價格因素 (Price)」、「其他因素 (Other)」分別標註 0 到 5。除了上述提及的欄位特殊處理，其餘的欄位若資料值為 Yes 或 No，則一律換成 1 與 0。之後將分類的欄位進行 one-hot encoding 處理。

3.2. Data Imputation

將資料先分成訓練資料與驗證資料後，先使用訓練資料進行插補方法的訓練，再分別擬合到訓練資料及驗證資料上。Khan and Hoque (2020) 提到 imputation 可分為 single imputation 及 multiple imputation[1]，scikit-learn 也提供兩種 imputation 的實作，包含 SimpleImputer、IterativeImputer 及 KNNImputer（此方法會在補值前先進行 scaling）。

[1] Khan, S.I., Hoque, A.S.M.L. (2020). SICE: an improved missing data imputation technique. *Journal of Big Data*, 7, 37. https://doi.org/10.1186/s40537-020-00313-w

SimpleImputer 提供填補缺失值的基本方法，缺失值可以使用平均、中位數、眾數來進行插補。IterativeImputer 將每個具有缺失值的特徵進行建模，一次指定一個特徵為 y，其餘的為 X 進行訓練，並以迭代的方式每次估計不同的特徵，直到每個有缺失值的特徵都差補完畢。KNNImputer 提供了基於 K-Nearest Neighbors 方法的缺失值插補。此概念是尋找尤拉距離最近的點，根據最近的 K 個點進行特徵加權運算，估計該缺失值。

我們在實驗中使用 SimpleImputer 與 IterativeImputer 進行補值，前者包含平均與中位數，後者則包含 Bayesian Ridge Regression、DecisionTree、Extra Trees Regression，此外，我們還採用全部補零，以及全部補零、加上 binary column 記錄是否缺值的做法。

在初步的驗證中[2]，我們發現以平均來說，採用何種補值方法並不會影響 F1-score太多；但針對此次大部分模型如 GradientBoosting 與 XGB，採用 Extra Trees Regression 的 multiple imputation 通常表現更佳。

3.3. Feature Scaling

本次分析嘗試無 scaling、 min-max scaling 與 z-score normalization，實驗顯示後者表現較佳[3]。

3.4. Feature Selection

本次分析皆使用 scikit-learn 的 Feature Selection方法，包含 VarianceThreshold、SelectKBest、LinearSVC、ExtraTreesClassifier。

VarianceThreshold 是一種基於變異數的特徵選擇法，若某特徵的變異數不滿足指定的閾值，則將該特徵剔除。預設的閾值為0.8。SelectKBest 可使用指定的單因子統計檢定方法進行運算，選出分數最高的k個特徵。預設方法為 chi square，k為 50。LinearSVC 是使用 SVM 的線性模型，並使用 L1 norm 來做 regularization 以避免模型發生 overfitting，來產生稀疏的估計結果，並選擇係數非零的特徵集合作為輸出的向量。其中，參數 C 控制稀疏性：C 越小，選擇的特徵越少。此方法預設的參數為 C=0.01, penalty="l1", dual=False。ExtraTreesClassifier 是一種基於決策樹的特徵選擇方法，使用 Bootstrap Aggregating 演算法，由多個決策樹的輸出投票決定出整個模型的輸出結果，可計算出特徵的重要性。預設參數 n_estimators 為 100。

經過多次調整後發現，使用預設值上下一倍的參數，交叉驗證的 F1-score其實沒有顯著的提升，因此後續皆使用原本的模型預設參數。經過實驗發現[4]，SelectKBest、LinearSVC 與 ExtraTreesClassifier表現優於 VarianceThreshold，其中又以 ExtraTreesClassifier 表現最佳。

---

[2] 使用 Training Method 中列出的 model，無 over-sampling 與 semi-supervise 步驟。

[3] 同樣使用 Training Method 中列出的 model，無 over-sampling 與 semi-supervise 步驟。

[4] 同樣使用 Training Method 中列出的 model，無 over-sampling 與 semi-supervise 步驟。

3.5. Over Sampling

由於資料集的 label 數量相當不平衡，我們加入 Over Sampling 步驟以期獲得更好的預測成效，本次分析的 Over Sampling 使用 Imbalanced Learn package 中的方法，包含 SMOTE、ADASYN、BorderlineSMOTE、SVMSMOTE。

SMOTE 是使用基本的差補法生成新樣本[5]。SMOTE 會隨機選擇所有可用的點來計算新樣本，平衡不同類別的樣本數量。ADASYN 同樣是使用差補法生成新樣本。與 SMOTE 不同的是，ADASYN 使用 K-Nearest Neighbors 分類器先選定 k 個鄰近的點來用於新樣本的生成。BorderlineSMOTE 是基於SMOTE 的改良演算法。首先將資料分成三類：(1) 雜訊 (2) 危險 (3) 安全。雜訊指的是所有鄰近的資料點皆屬於不同的類別；危險指一半的鄰近資料點屬於不同的類別；安全指所有鄰近的資料點皆屬於相同的類別。BorderlineSMOTE 會抽取屬於危險類別的資料進行新樣本的生成。SVMSMOTE 是使用 SVM 分類器來查找支持向量並一此生成新樣本。SVM 分類器的 C 參數可調整支持向量的多寡。

而經過實驗發現[6]，SMOTE、BorderlineSMOTE 與 SVMSMOTE 三者表現較佳。

3.6. Training Method

以下列出本研究採用的模型，其中 Decision Tree、Random Forests、XGBoost、GBM 等模型也會加入半監督式學習進行模型訓練。

- **Logistic Regression (penalty='l2')**

  其對線性關係的擬合效果極佳、計算快，對於模型的解釋也較為容易，但容易有過擬合的問題。

- **Decision Tree**

  Decion Tree 為一樹狀結構，由樹根開始每經過一個 node 就將資料分割到不同邊，分割的原則是：這樣的分割要能得到最大的資訊增益 (Information gain，簡稱 IG)。常見的資訊量有兩種：Entropy 以及 Gini Impurity，而資訊增益為分割前後資訊量的差值。資訊量越低，資料的同質性就越高，因此分類效果越好。
  Decion Tree 的訓練速度快，且分類過程非常直觀，因此具解釋性。

- **Random Forests (n_estimators=100)**

  Random Forests 是一個包含多個決策樹的分類器[7]，使用 Bootstrap Aggregating 演算法，由多個決策樹的輸出投票決定出整個模型的輸出結果。Random Forests 中的每棵樹依照一個獨立隨機抽樣的向量產生，而森林中的每一棵樹皆從相同的分布中產生。Random Forests 的泛化誤差的收斂會受到樹的數量的限制。此外，Random Forests 的泛化誤差也會受到裡面每棵樹的分類器的強度及樹之間的相關性影響。

---

[5] Over-sampling, https://imbalanced-learn.org/stable/over_sampling.html

[6] 使用 Training Method 中列出的 model，無 semi-supervise 步驟。

[7] Breiman L. (2001). Random Forests, *Machine Learning*, 45(1): 5–32. doi:10.1023/A:1010933404324

從訓練資料中產生多棵決策樹，之後進行投票，時間複雜度取決於決策樹的數量、資料維度及訓練資料集的大小等，樹的部分大多採用二元樹，與其他機器學習方法相比，訓練速度快，並且可處理大量特徵、穩定性、受離群值的影響不大。以外，可以透過 Random Forests 觀察每一個特徵的重要程度，因此模型具解釋性。

- **GBM (n_estimators=100)**

  GBM 全名為 Gradient Boosting Machine，也是一種 boosting 的技巧，scikit-learn 中使用 regression tree 作為弱預測模型。特點在於 GBM 使得更多 loss function 可微分，使模型能處理 Regression、Classification、Ranking 等不同的問題，增加 model 的通用性。此模型利用 Stochastic Gradient Boosting、分支樣本數下限、樹複雜度的處罰權重來減少 overfitting 的可能性。GBM 通常可以達到更佳的預測效果，但由於模型複雜，可讀性不高，需要的計算能力較大，應用上也較為困難。

- **XGBoost (n_estimators=100, learning_rate=0.3, use_label_encoder=False)**

  XGBoost 是 Extreme Gradient Boosting 的縮寫[8] [9]，是基於 Gradient Boosting 架構下的監督式機器學習演算法，提供 GPU 平行處理，使效率及準確度提升。XGboost ，和 Random Forests 一樣，是採取特徵隨機抽樣的機制，隨機抽取特徵來生成決策樹，所以不會每一次都使用全部的特徵來參與決策。XGboost 中每一棵樹是互相關聯的，希望後面生成的樹能夠修正前面一棵樹犯錯的地方。另外，XGboost 加入了 regularization 的機制，避免過度受到雜訊的影響而發生 overfitting。
  XGboost 具有極高的可擴展性，只要機器的 RAM 還有空間就可以進行訓練。

- **DNN (learning_rate= 0.003,weight_decay =0.0001, drop_out=0.4)**

  DNN 是 Deep Neural Network 的縮寫，是由若干神經元所組成的網路，為分層的結構，每層由若干個神經元所組成，每個神經元為一個分類器 (perceptron)。透過調整層數以及每層的神經元數量，理論上能逼近任何數學模型，即 hypothesis set 夠大，能夠在不同的 dataset 上都有不錯的訓練效果；然而參數較多，因此效能較差，並且需要大量的data來訓練，也容易出現 overfitting 的狀態。
  為了解決 overfitting 的問題，我們在 model 中加入 weight decay 和 drop out 來緩解。weight decay 即為 regularization，而 drop out 是在訓練過程中，以一定的機率排除某個神經元（輸出為 0），來防止 model overfit training data。

---

[8] Tianqi Chen and Carlos Guestrin. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, 785–794. DOI:https://doi.org/10.1145/2939672.2939785

[9] Introduction to Boosted Trees, https://xgboost.readthedocs.io/en/stable/tutorials/model.html

4. Evaluation

Model	Validation Score	Public Score	Private Score
XGB	0.33134	0.29947	0.31902
XGB (SVMSMOTE)	0.33364	0.31590	0.28971
GBM (SMOTE)	0.33404	0.31463	0.29990
XGB (Semi)	0.22513	0.30644	0.30052
GBM (Semi)	0.24266	0.35909	0.29407
XGB (SMOTE, Semi)	0.22967	0.33440	0.29949
GBM (SMOTE, Semi)	0.24334	0.34347	0.30027

我們從上百種排列組合中，依據是否有做 oversampling、是否有 semi-supervised 選出表現較佳的組合，可以看到有做 semi-supervised 的模型在 validation 的表現較差，但加入 test set 做訓練後，在 Kaggle 上面的 public score 與其他模型相差無幾，private score 則是表現得更好，這可能是因為 sub-train set 的資料量較少；oversampling 對於 XGB (Semi) 較有正面影響，對於其他模型則較無效果。

5. Conclusion

綜上所述，我們認為使用 Extra Trees 的 multiple imputation 與 feature selection 後，利用 GBM 做 semi-supervised training，可以達到最佳的預測效果，對於未來的 unlabeled data 也能加入進行訓練；但相對其他 model 來說，需要較高的運算能力，interpretability 也較弱。over sampling 雖然可以做出假樣本使得 label balance，但仍須更多的調整與測試，避免做出來的假樣本影響 model 做出錯誤的判斷。

6. Workload Assignment

- R10725023 吳琦艾：Logistic Regression、GBM 模型訓練
- R10725045 蔡祐琳：XGBoost、Random Forests 模型訓練
- R10922078 賴侃軒：Decision Tree、DNN 模型訓練
- 共同部分：實作 Data Preprocessing、Data Imputation、Over Sampling、書面報告撰寫