Kaggle 個人競賽實驗報告 Kaggle Individual Competition Report

張哲安/M946113003/大數據科技及管理研究所

一、資料集前處理 Dataset Preprocessing

麻醉方式中相同的合併一起。手術名稱、藥物使用、身上管路及擺放位置中相似的內容，組合起來並分別以每加總至 250、1300、760 進行類別分組，創造出每組大致相同數量的組別，如果是空值或沒有出現在資料中的放入最後一類。藥物使用中，找出使用抗凝血劑、心律不整、心衰竭者。找出使用高血壓、糖尿病者。找出使用癲癇藥物。找出使用止吐藥物。找出使用嗎啡類止痛藥物，建立二元欄位。以上述建立一個 0-8 藥物組合使用的類別。手術名稱中，以開刀的部位進行 0-9 的分類，分別為心、腦、胸肺、腹、婦、脊髓、耳鼻喉、下肢、上肢及眼。身上管路中，計算身上總管路數。檢驗數值中，建立 Troponin I,NA,K,WBC,HB,PLT,INR,CREA,O,GLU,CRP,ALB 是否為異常的二元欄位，並以 CREA、年齡、性別計算是否有 ESRD。抽血欄位建立是否所有為正常值的二元欄位。年齡建立 5 類的年齡分組。以建立的二元藥物欄位，建立疾病總和欄位。把加護病人、麻醉方式、病人來源轉換成數字類別，空值建立獨立一類。對類別資料進行 Targetencoder，再把所有進行標準化。身高、體重空值使用 KNN，K 設為 5 來填補。填補後計算 BMI。並建立 3 類組合。

二、系統架構及方法說明

在特徵整理過後，過程使用 StratifiedKfold 切 10 等份。模型訓練過程中因不平衡資料使用 Balaced 權重調整。使用 BayesSearchCV 調參找 250 組參數。用調參後的 XGBoost 進行特徵重要性篩選。保留前 80%重要特徵進行其他模型的訓練。模型使用包括 Naïve Bayes、Logistic Regression、Decision Tree、KNN、SVM、XGBoost、Random Forest、LightGBM、Catboost 及 StackClassifer，在 Stack 中，基礎模型使用 XGBoost、Random Forest、LightGBM 及 Catboost，並於最終模型使用 Logistic Regression。

三、實驗記錄 Experiments

1.Machine Learning Classifiers

Method	Training Data k-fold Average-Micro Precision/Recall/F1	Training Data k-fold Average-Macro Precision/Recall/F1	Test Data Public Leaderboard Score/Rank/filename	Test Data Public Leaderboard Score/Rank/filename	Parameters for training model
Naïve Bayes#1	0.458/0.458/0.458	0.343/0.368/0.312	0.311/kaggle_submission	0.297	NA
Naïve Bayes#2	0.449/0.449/0.449	0.339/0.372/0.314	0.319/kaggle_submissionNB01	0.297	NA
Naïve Bayes#3	0.444/0.444/0.444	0.34/0.379/0.32	0.328/kaggle_submissionNB02	0.307	NA
Logistic Regression #1	0.61/0.61/0.61	0.597/0.604/0.597	0.452/1101Logistic1	0.424	NA
Logistic Regression #2	0.776/0.776/0.776	0.701/0.61/0.626	0.305/1101Logistic2	0.288	NA
Logistic Regression #3	0.476/0.476/0.476	0.442/0.579/0.45	0.457/1102Logistic3	0.426	NA
Decision Tree #1	0.506/0.506/0.506	0.398/0.398/0.397	0.387/1031DT01	0.394	NA
Decision Tree #2	0.44/0.44/0.44	0.414/0.552/0.415	0.419/1031DT02	0.394	criterion:gini,entropy , max_depth:3-10 , max_features:1-20 , min_samples_leaf:5-50 , min_samples_split:10-100
Decision Tree #3	0.45/0.45/0.45	0.417/0.547/0.422	0.426/1031DT03	0.416	同上
KNN #1	0.568/0.568/0.568	0.506/0.456/0.467	0.437/1102KNN01	0.445	NA

KNN #2	0.6/0.6/0.6	0.565/0.476/0.5	0.456/1102KNN02	0.447	n_neighbors:1-20 , weights:uniform,distance , algorithm:auto, ball_tree, kd_tree, brute , leaf_size:20-50 、 p:1-2
KNN #3	0.58/0.58/0.58	0.537/0.422/0.447	0.441/1104KNN3	0.433	同上
SVM #1	0.625/0.625/0.625	0.48/0.399/0.408	0.414/1105SVC01	0.401	NA
SVM #2	0.628/0.628/0.628	0.478/0.356/0.408	0.414/1211SVC02	0.401	NA
SVM #3	0.627/0.627/0.627	0.481/0.354/0.408	0.414/1213SVC03	0.396	C:0.1-100 , gamma:0.001-10 , kernel:rbf,poly , degree:2-4

2.你的方法 Your Methods

Method	Training Data k-fold Average-Micro Precision/Recall/F1	Training Data k-fold Average-Macro Precision/Recall/F1	Test Data Public Leaderboard Score/Rank/filename	Test Data Public Leaderboard Score/Rank/filename	Parameters for training model
XGBoost#1	0.619/0.619/0.619	0.554/0.421/0.445	0.507/1105XGB01	0.481	learning_rate:0.01-0.2 , max_depth:3-7 , subsample:0.7-1 , colsample_bytree:0.7-1 , n_estimators:100-1500 , min_child_weight:1-5
XGBoost#2	0.589/0.619/0.619	0.584/0.423/0.449	0.516/1107XGB2	0.498	同上
XGBoost#3	0.624/0.624/0.624	0.584/0.425/0.449	0.510/1107XGB3	0.502	同上
XGBoost#4	0.628/0.628/0.628	0.589/0.431/0.459	0.52/1116XGB4	0.499	learning_rate:0.01-0.1 , max_depth:3-8 , subsample:0.6-1 , colsample_bytree:0.6-1 , n_estimators:100-1000 , min_child_weight:1-7 , gamma:0-3

Catboost#1	0.617/0.614/0.614	0.532/0.547/0.538	0.536/1203catboost	0.548	learning_rate:0.001-0.05,iterations:1000-3000,depth:4-7,min_data_in_leaf:10-50,leaf_estimation_iterations:5-10,grow_policy:SymmetricTree,Depthwise,bootstrap_type:Bernoulli,subsample:0.6-0.8,l2_leaf_reg:3.0-10.0
Catboost#2	0.602/0.602/0.602	0.523/0.605/0.548	0.54/1213catboost1	0.542	NA
Catboost#3	0.62/0.62/0.62	0.527/0.606/0.552	0.537/1213catboost2	0.551	NA

4.討論 Discussion

如何從原始資料中建立對模型有用不稀疏的特徵極為重要，並且了解不同模型的優缺點。Catboost 模型有一些特性較適合此資料及我前處理的方式，類別項目多、空值及文本。此模型可以轉換類別特徵至數值型態。並且使用 Ordered Target Encoder，相較於一般常用的 Target Encoder 有資料洩漏疑慮，造成泛化能力不好。此模型方式是對資料進行排序，只對 n-1 筆資料進行數值計算轉換，減少了資料洩漏的問題，提高了泛化的能力。還有文本特徵透過 Bag of words 或其他方式轉換至數值型態。模型生成多個對稱樹進行 Boosting，相較其他 Boosting 使用非對稱樹更弱的弱分類器，而解決過擬合提高泛化能力的其中一個方式就是簡化模型，這可能也是可以提高泛化能力的原因。所以相對於 XGBoost 在對 Kaggle 資料預測時，可以有好一點的成績。

5.外部資源與參考文獻

[1] Governance, "Statement on ASA Physical Status Classification System," American Society of Anesthesiologists (ASA), https://www.asahq.org/standards-and-practice-parameters/statement-on-asa-physical-status-classification-system , Dec. 2020 (accessed Nov. 2024).

[2] Yandex, "Transforming text features to numerical features," Catboost, https://catboost.ai/docs/en/concepts/algorithm-main-stages_text-to-numeric , (accessed Dec. 2024).

[3] Yandex, "Transforming categorical features to numerical features," Catboost, https://catboost.ai/docs/en/concepts/algorithm-main-stages_cat-to-numberic , (accessed Dec. 2024).

[4] Yandex, "Feature importances," Catboost, https://catboost.ai/docs/en/features/feature-importances-calculation , (accessed Dec. 2024).

[5] Yandex, "Parameter tuning," Catboost, https://catboost.ai/docs/en/concepts/parameter-tuning , (accessed Dec. 2024).