Machine Learning Homework2

# Introduction

此篇文章為機器學期課程的第二份作業，使用 Kaggle 上的 Car Insurance Claim Prediction 資料集進行多種分類任務與 cross-validation 的實作，並比較多種調整方式與已被廣泛使用的分類套件之間的成效差異。

# Methodology of Use

## Naïve Bayer classifier

## Random Forest Classifier

## Random Forest Classifier (sklearn)

## XGBoost

## Catboost

## LightGBM

# Experiment

## Preprocessing

各欄位皆沒有缺失值，將資料集以 80:20 比例切分 train/test，對類別欄位和數值欄位分別觀察與前處理，若類別小於 10 種的欄位採取 one-hot encoding 的方法處理；若類別大於 10 種的欄位則使用 target-encoding。

## Experiment settings

先不採取 cross-validation 的流程測試模型成效，接著再套用 k = 3, 5, 10 的 K-Fold 模式來比較模型，由於資料非常的 unbalanced ，label 為 1 的僅佔 93%，所以選擇使用 f1-score 查看模型好壞。

## Performance on classification task

| Method | F1 |
| --- | --- |
| Naïve Bayer classifier | 0.120 |
| Random Forest Classifier | X |
| Random Forest Classifier (sklearn) | 0.162 |
| XGBoost | 0.138 |
| CatBoost | 0.158 |
| LightGBM | 0.160 |

## Cross-Validation on classification task

| Method \ k | 3 | 5 | 10 |
| --- | --- | --- | --- |
| Naïve Bayer classifier | 0.128 | 0.130 | 0.105 |
| Random Forest Classifier | X | X | X |
| Random Forest Classifier (sklearn) | 0.164 | 0.166 | 0.141 |
| XGBoost | 0.140 | 0.152 | 0.115 |
| CatBoost | 0.153 | 0.161 | 0.132 |
| LightGBM | 0.169 | 0.168 | 0.122 |

# Conclusion

在不套用 CV 的模型評估中，常常會因為資料切分的隨機性，而造成同樣模型架構，遇到不同資料導致的成效不同，從 CV 的表格中就可以看出，在 k = 3 或 5 的情形下，都是由 LightGBM 勝出，但是在 k = 10 的時候，Random Forest 模型所下降的成效較低，遇見不同資料時，Random Forest 的通用性就展現出來了。

以上所有模型的相關實作程式碼都有上傳至 Github：

https://github.com/littlemilkwu/ds\_ncku\_machine\_learning