客戶續約金額預測

設計文件

隊伍: 都看看是誰來了 (第 8 名)

成員：陳俊穎、莊博勝、吳亦振

摘要

我們這次使用的訓練模型為 LightGBM，資料經過轉換以及 One-Hot encoding 和 feature engineering 後，訓練資料集初步有 1499 個特徵。我們主要建立兩個模型來預測客戶續約金額，第一個模型為預測該客戶的續約金額成長率；第二個模型則預測該客戶是否續約（1 代表續約），我們利用第二個模型的預測結果對第一個模型的預測結果做修正，當作客戶續約金額的預測。

環境

OS：WIN 10

CPU：Intel i5 - 8500

RAM：16GB

程式語言：Spyder (Python 3.6)

函式庫：

numpy
pandas
datetime
lightgbm

sklearn.cross_validation
sklearn.preprocessing
sklearn.metrics

特徵

我們的訓練集使用 Policy_0702 與 Claim_0702 大多數的特徵，除了 Claim_0702 中的 Policy_Number、Claim_Number、Vehicle_identifier ，以及 Policy_0702 中的 Policy_Number、Insured's_ID、Prior_Policy_Number、Vehicle_identifier、Vehicle_Make_and_Model2、Coding_of_Vehicle_Branding_&_Type、aassured_zip。移除上述特徵的原則有二：(1) 主觀認定該特徵無資訊，像是 ID 類型的。 (2) 該類別型特徵中最高比例的種類小於有效樣本數的 5 % 則代表該類別型特徵下每個種類的樣本太少。移除上述特徵後，增加 Policy_0702 中的類別型特徵對 Premium 做 mean encode，以及 Policy_0702 中兩兩交互作用的類別型特徵。

最後經過 One – hot encode 與其他處理方法 (Describtion.txt) ，初始訓練集共有 1499 個特徵。

訓練模型

對於「續約金額成長率」以及「是否續約」的預測，我們皆使用 LightGBM 機器學習模型，使用參數則參考參賽者 ***diffusion*** 在討論區提供的參數。

模型一 (續約金額成長率) LightGBM 參數：

```
param = {
    'boosting_type': 'gbdt',
    'class_weight': None,
    'colsample_bytree': 0.733333,
    'learning_rate': 0.00764107,
    'max_depth': -1,
    'min_child_samples': 460,
    'min_child_weight': 0.001,
    'min_split_gain': 0.0,
    'n_estimators': 3000,
    'n_jobs': -1,
    'num_leaves': 77,
    'objective':None,
    'random_state': 42,
    'reg_alpha': 0.877551,
    'reg_lambda': 0.204082,
    'silent': True,
    'subsample': 0.949495,
    'subsample_for_bin': 240000,
    'subsample_freq': 1,
    'metric': 'l1'
    }
```

模型二 (是否續約) LightGBM 參數：

```
param = {
    'boosting_type': 'gbdt',
    'class_weight': None,
    'colsample_bytree': 0.733333,
    'learning_rate': 0.00764107,
    'max_depth': -1,
    'min_child_samples': 460,
    'min_child_weight': 0.001,
    'min_split_gain': 0.0,
    'n_estimators': 3000,
    'n_jobs': -1,
    'num_leaves': 20,
    'objective':'binary',
    'random_state': 42,
    'reg_alpha': 0.877551,
    'reg_lambda': 0.204082,
    'silent': True,
    'subsample': 0.949495,
    'subsample_for_bin': 240000,
    'subsample_freq': 1,
    'metric': 'auc'}
```

訓練方式及原始碼

針對「續約金額成長率」的預測，我們不加入成長率大於 (Q3 + 0.5 * IQR) 和下年度無續約的樣本進入 CV 的訓練集。訓練方式為 5 - fold CV，每次建模都會得出無重要性的特徵，將這些特徵取交集後移除，再進入下一次訓練，重複以上動作。如果比前一次的訓練結果表現還差則停止循環，並以前一次的訓練為準。每次建模都對目標預測對象(testing - set) 預測，得出 5 次預測值後再取平均，得出第一個模型的預測結果。

對於「是否續約」的訓練方式與上述相同，差異在於不移除特定樣本。得出 Validate – set 的續約機率後，我們找到最好的 threshold 使得預測準確度最高，再以此 threshold 對目標預測對象的預測機率做分類，得出第二個模型的分類預測結果。

提交的結果為：

Prediction = (1 + 預測成長率) * 上次保費續約總額

Prediction [分類預測結果 == 0] = 0

程式碼：

https://drive.google.com/open?id=1BdVwbxSBbaPnucOHtTcriEAexm-qLR2T

結論

1. 對於 EDA 的觀察沒有花太多時間深入研究，導致一直無法突破瓶頸。
2. 在預測是否續約時，只能成功預測到 25%沒續約的人，是這次競賽最令人頭痛的部份，尤其是沒預測到大客戶不續約。考慮到出險紀錄會導致隔年保費上升，影響客戶續約意願，因此有嘗試將成長率的預測加入分類模型的特徵，但分類情況沒什麼改善。
3. 發現出險客戶的成長率預測非常差，沒出險客戶的成長率預測得還不錯，嘗試將這兩群分開建模預測，但沒有顯著的效果。
4. 先前續約金額較多的客戶(Total_Premium > 10000) 的 MAE 很大，小客戶的 MAE 則是 1000 以下，同時也發現大戶和出險情況有些關聯。
5. 因為 LightGBM 是很厲害的模型，所以就算用兩層的 Stack model 也不會有太多的進步，也有嘗試 NN，但表現比 LightGBM、XGBoost 還要差。