Báo cáo Solution cho bài Credit Scoring

1. **Phân tích dữ liệu:**
   - Số lượng dữ liệu là ít: 30k dòng cho tập train và 20k dòng cho tập test.
   - Dữ liệu thuộc loại imbalance data: tỉ lệ nhãn 1 / nhãn 0 xấp xỉ 1.6%
   - Hầu hết các trường dữ liệu đã được mã hóa nhằm bảo mật thông tin user.
   - Số lượng missing value là tương đối nhiều và hầu hết các trường dữ liệu đều có missing value => cần phải cleaning data tốt để improve score.
   - Một vài trường dữ liệu cần lưu ý:

   FIELD_3: dải giá trị của trường này có chu kỳ 365, nghĩa là giá trị tập trung ở các số 1, 2, 3, … sau đó nhảy lên 366, 367, …, 730, 731 => rất có thẻ đây là encode của một biến thời gian nào đó

   Age_source1 và age_source có sự sai khác nhau, có thể 1 trong 2 bị missing hoặc khác nhau.

   FIELD_7: data type của dữ liệu ở dạng list

   FIELD_9: sau khi search dãy dữ liệu của trường này thì phát hiện đây là các mã tương ứng đối tượng tham gia bảo hiểm y tế.

   Tương tự thì FIELD_39 là mã quốc gia.
2. **Data processing**
   - Với các trường dữ liệu dạng numberic, nếu bị missing thì sẽ được fill thành -99, nếu có giá trị ‘nan’ hoặc ‘None’ thì được replace thành -1.
   - Tương tự với các trường dữ liệu dạng object, nếu bị missing thì sẽ được fill thành ‘Missing’.
   - Dữ liệu dạng text như ‘province’, ‘district’, ‘maCv’ sẽ được chuẩn hóa về viết thường và được tokenize.
   - Dữ liệu dạng có thể order như FIELD_41, FILED_42 sẽ được map thành các số tương ứng (I : 1, II : 2, III : 3, IV : 4)
3. **Feature engineering**
   - MaCv và FIELD_7 sẽ được convert về dạng text và sẽ được Tf-idf transform sang dạng vector rồi áp dụng thuật toán giảm chiều SVD để cho ra vector 2 chiều.
   - Bổ sung thêm các feature dạng Frequency Encoding của các dữ liệu dạng object/category.
   - Thêm 1 feature biểu thị sự khác nhau giữa age_source1 và age_source2.
   - Thêm 1 feature là phần dư của FIELD_7 chia cho 365.

- Tiến hành sinh các feature mới bằng cách sử dụng WEIGHT OF EVIDENCE: https://medium.com/@sundarstyles89/weight-of-evidence-and-information-value-using-python-6f05072e83eb

4. **Training model**

   Các bước training như sau:
   - Cross validate dữ liệu theo phương pháp k-fold với k = 5.
   - Training với thuật toán Gradient Decsent Boosting Tree (thư viện LightGBM)
   - Chạy Bayes Opimize Hyperparameters để tìm bộ parameter tối ưu nhất.
   - Training với bộ params tối ưu.

5. **Kết quả**

   Average gini score trên tập validate của các fold đạt **0.252.**

   Quá trình improve kết quả:

   + Ban đầu chỉ chạy mô hình baseline (chưa preprocessing, chưa feature engineering), kết quả đạt 0.20.

   + Sau quá trình preprocessing, kết quả tang lên 0.23.

   + Feature engineering và chạy bayes optimize, kết quả tang lên 0.24.

   + Bổ sung thêm các feature mới từ phương pháp WoE, kết quả tang lên 0.252.