#	Trường dữ liệu	Kiểu dữ liệu	Tỷ lệ missing	Ý nghĩa trường (Dự đoán)	Nhận xét (Lỗi, phân bố)	Phương án xử lý & chuẩn hoá (Cách thức chuẩn hoá, sửa lỗi)
1	FIELD_1	numerical	0.00%		dữ liệu 0,1	Giữ nguyên kiểu categorical
2	FIELD_10	categorical	32.25%		Gồm dữ liệu: T1, GH, None	xử lý dữ liệu None, Missing
3	FIELD_11	categorical	32.36%		Tường dữ liệu số từ 0 đến 69 và None	xử lý dữ liệu None, Missing
4	FIELD_12	categorical	32.25%	Liên quan đến bảo hiểm y tế, có thể là bảo hiểm y tế đã được hưởng	Dữ liệu train: 0,1,HT,TN, None (người được hưởng lương hưu, người trợ cấp thất nghiệp) dư liệu test: 0,1, DK,GD, XK, DN, DT, HT (người sinh sống tại vùng khó khăn, đóng bảo hiểm theo gia đình, người làm doanh nghiệp, người dân tộc thiểu số, người hưởng lương hưu) - Các mã này xuất hiện rất ít, chủ yếu là 0,1, None	xử lý dữ liệu None, Missing và noise
5	FIELD_13	categorical	32.33%		- Gồm dữ liệu 2 kí tự và số 0,4,8,12 - Những giá trị tệp test ko có trong train: A1, AE,BJ, BW,CG,CL,DL,DM,EK,EU,FT,H2,H5,H7,IS,KX,N7, NM,NP,NY,QQ,QS,SB,SZ,ZA	
6	FIELD_14	numerical	0.00%		Giống FIELD_15 đến 98%	
7	FIELD_15	numerical	0.00%		Encode của FIELD_52. If FIELD_52 == 30.955 then 0 else 1	
8	FIELD_16	numerical	32.36%	Các trường từ 16 - 26 nếu một trường có giá trị thì các trường còn lại cũng sẽ có và nếu để trống sẽ là tất cả cùng trống => khả năng đây là một chuỗi dữ liệu nào đó liên quan đến nhau được chia nhỏ ra	Phân bố các giá trị giữa 2 tập train, test giống nhau	
9	FIELD_17	categorical	32.36%	Các giá trị xuất hiện G2->GX (G10?). Khả năng rank theo cái gì đó nên ngoài lựa chọn là để categorical thì có thêm lựa chọn là chuyển về dạng số xét theo numeric feature.	Các giá trị giữa train, test không match nhau	Giữ nguyên kiểu categorical
10	FIELD_18	categorical	32.36%		Phân bố các giá trị giữa 2 tập train, test giống nhau và chỉ có 2 giá trị True và False	Giữ nguyên kiểu categorical
11	FIELD_19	categorical	32.36%			Giữ nguyên kiểu categorical
12	FIELD_2	numerical	1.58%		dữ liệu 0,1	Giữ nguyên kiểu categorical
13	FIELD_20	categorical	32.36%			Giữ nguyên kiểu categorical
14	FIELD_21	numerical	32.36%		Phân bố các giá trị giữa 2 tập train, test giống nhau bao gồm (0,1,2) trong đó giá trị một chiếm đến 99,99%	
15	FIELD_22	numerical	32.36%		Dữ liệu numeric nhưng giá trị trong 2 tập train test không match nhau	

#	Trường dữ liệu	Kiểu dữ liệu	Tỷ lệ missing	Ý nghĩa trường (Dự đoán)	Nhận xét (Lỗi, phân bố)	Phương án xử lý & chuẩn hoá (Cách thức chuẩn hoá, sửa lỗi)
16	FIELD_23	categorical	32.36%	Trường này bằng kiểm tra dữ liệu missing ở các trường khác, nếu TRUE là có dữ liệu, missing là ở các trường khác có thể là 1 kiểm tra 1 data khác có dữ liệu hay không	Chỉ có duy nhất một giá trị True	Giữ nguyên kiểu categorical
17	FIELD_24	categorical	32.36%	Mã vùng sinh sống của người tham gia BHYT (K1,K2,K3)	Phân bố các giá trị giữa 2 tập train, test giống nhau, trong đó chủ yếu xuất hiện các giá trị None	Giữ nguyên kiểu categorical
18	FIELD_25	categorical	32.36%		Phân bố các giá trị giữa 2 tập train, test giống nhau và chỉ có 2 giá trị True và False	Giữ nguyên kiểu categorical
19	FIELD_26	categorical	32.36%			Giữ nguyên kiểu categorical
20	FIELD_27	categorical	32.36%		- Dự liệu: TRUE, FALSE - woe_top: 106 độ quan trọng thấp - 100% values appear in both train & test.	- Dữ liệu các trường này về cơ bản đều là dạng TRUE, FALSE nên không được xử lý gì đặc biệt, nên đưa về dạng categorical features. Side notes: một bài chia sẻ trên forum có dùng thêm one-hot encoding với các trường: ['FIELD_8', 'FIELD_10', 'FIELD_17', 'FIELD_24', 'FIELD_35', 'FIELD_41', 'FIELD_43', 'FIELD_44'] See [1] https://forum.machinelearningcoban.com/t/kalapa-s-credit-scoring-challenge-13-solution-02-02-0-2729-gini-score/7139.
21	FIELD_28	categorical	32.36%		- Dự liệu: TRUE, FALSE - woe_top: 102 -> độ quan trọng thấp - 100% values appear in both train & test.	
22	FIELD_29	categorical	32.36%		- Dự liệu: TRUE, FALSE, None - woe_top: 151 - 4 unique values in train, 3 values in test.	
23	FIELD_3	numerical	1.58%	Ngày tháng	Khi lấy unique thì nhận thấy các giá trị không liên tục mà chia ra thành các cluster. Trong 1 cluster giá trị liên tục (thường cách nhau 1) và không có quá 31 giá trị. Từ đó có thể đoán được 1 cluster là 1 tháng	
24	FIELD_30	categorical	32.36%		- Dự liệu: TRUE, FALSE, None - woe_top: 7 - 100% values appear in both train & test.	
25	FIELD_31	categorical	32.36%		- Dự liệu: FALSE, None - woe_top: 38 - 100% values appear in both train & test.	
26	FIELD_32	categorical	0.00%		- Dữ liệu : 0,1 - Desc: woe-top: 103 - Solution: drop - 100% values appear in both train & test.	
27	FIELD_33	categorical	0.00%		- Dữ liệu : 0,1 - woe_top: 116 - 100% values appear in both train & test.	

#	Trường dữ liệu	Kiểu dữ liệu	Tỷ lệ missing	Ý nghĩa trường (Dự đoán)	Nhận xét (Lỗi, phân bố)	Phương án xử lý & chuẩn hoá (Cách thức chuẩn hoá, sửa lỗi)
28	FIELD_34	categorical	0.00%		- Dữ liệu : 0,1 - woe_top: 177 - 100% values appear in both train & test.	
29	FIELD_35	categorical	32.36%		- Dữ liệu: Zezo, One, Two, Three, Four - woe_top: 13 - 100% values appear in both train & test.	
30	FIELD_36	categorical	32.36%		- Dữ liệu: TRUE, FALSE, None - woe_top: 84 - 4 unique values in train, 6 unique values in test	
31	FIELD_37	categorical	32.36%		- Dữ liệu: TRUE, FALSE, None - woe_top: 109 - 100% values appear in both train & test.	
32	FIELD_38	categorical	32.36%		- Dữ liệu: TRUE, FALSE - woe_top: 134 - 100% values appear in both train & test.	
33	FIELD_39	categorical	32.36%	Quốc tịch người tham gia BHYT	Phân bố các giá trị khá giống nhau trên cả train/test (VN chiếm ~ 35%, None ~ 30%, các giá trị khác từ 0-1%), Có 6 giá trị trong test không có trong train	Giữ nguyên kiểu categorical, loại bỏ 6 giá trị
34	FIELD_4	numerical	1.58%		dữ liệu số từ 0 đến 12	Giữ nguyên kiểu categorical
35	FIELD_40	categorical	32.36%		Phân bố giá trị tương đối giống nhau trên cả train/test (None chiếm ~ 50%, giá trị 1 chiến khoảng 16%). Có những giá trị kiểu chuỗi lạ, vd: 02 05 08 11, 08 02, 05 08 11 02	Giữ nguyên kiểu categorical
36	FIELD_41	categorical	32.36%	Mức hưởng BHYT?	Phân bố các giá trị khá giống nhau trên cả train/test, nhiều nhất là mức 1 chiếm khoảng ~43%. Giá trị None khả năng là dữ liệu ngoại lai	Giữ nguyên kiểu categorical
37	FIELD_42	Categorical	32.36%		Phân bố các giá trị khá giống nhau trên cả train/test, giá trị Zero chiếm khoảng ~67%. Giá trị None khả năng là dữ liệu ngoại lai	Giữ nguyên kiểu categorical
38	FIELD_43	categorical	32.36%		Phân bố các giá trị khá giống nhau trên cả train/test, giá trị None chiếm khoảng ~64%, còn lại 0-2%	Giữ nguyên kiểu categorical
39	FIELD_44	categorical	32.36%		Phân bố các giá trị khá giống nhau trên cả train/test, giá trị ↔'One' chiếm khoảng ~49%, 'Two' chiếm ~ 18%, None ~ 0.1%.	Giữ nguyên kiểu categorical
40	FIELD_45	categorical	32.36%		Phân bố các giá trị khá giống nhau trên cả train/test, giá trị ↔'1' chiếm khoảng ~52%, '2' chiếm ~ 15%, None và '3' ~ 0.02%.	Giữ nguyên kiểu categorical

#	Trường dữ liệu	Kiểu dữ liệu	Tỷ lệ missing	Ý nghĩa trường (Dự đoán)	Nhận xét (Lỗi, phân bố)	Phương án xử lý & chuẩn hoá (Cách thức chuẩn hoá, sửa lỗi)
41	FIELD_46	categorical	0.00%		Phân bố các giá trị khá giống nhau trên train/test, chỉ có 2 giá trị 0 chiếm ~60%, 1 chiếm khoảng 40%	Giữ nguyên kiểu categorical
42	FIELD_47	categorical	0.00%		Phân bố các giá trị khá giống nhau trên train/test, chỉ có 2 giá trị 'True' chiếm ~58%, 'False' chiếm khoảng 42%	Giữ nguyên kiểu categorical
43	FIELD_48	categorical	0.00%		Phân bố các giá trị khá giống nhau trên train/test, chỉ có 2 giá trị '↔True'↔ chiếm ~69%, 'False' chiếm khoảng 31%	Giữ nguyên kiểu categorical
44	FIELD_49	categorical	0.00%		Phân bố các giá trị khá giống nhau trên train/test, chỉ có 2 giá trị '↔True'↔ chiếm ~88%, 'False' chiếm khoảng 12%	Giữ nguyên kiểu categorical
45	FIELD_5	numerical	1.58%	Có liên quan đến Field 7 nhưng chưa rõ ý nghĩa	Giá trị discrete từ 0 đến 14 và nan. Số lượng count giảm dần từ 0 đến 14 trong train và test (trừ trường hợp giá trị 1 và 2). Field 5 có thể liên quan đến Field 7. Chuỗi dài nhất của Field 7 là 14 phần tử, số lớn nhất của Field 5 là 14. Ngoài ra Field 5 có vẻ tỉ lệ với Field 7, ví dụ Field 7 là [] thì Field 5 là 0, Field 7 có 6 phần tử thì Field 5 khoảng 6	
46	FIELD_50	numerical	32.36%	Encode từ data liên quan đến thu nhập hay tiền đóng bảo hiểm? FIELD_50 đến 53 có thể liên quan với nhau theo thời gian	- Giá trị phổ biến nhất (29.770) là có vấn đề. Khi gộp cả train và test thì phát hiện ra giá trị này xuất hiện đúng 29669 lần (tức là bằng 29.770x1000-1) - Các giá trị còn lại đều nằm trong khoảng 60~67 - 1 vài giá trị luôn luôn có nhãn là GOOD: 67.477, 67.61, 67.634	các FIELD từ 50 đến 57 có thể là quan trọng vì nó có tận 3 cách encode khác nhau cho cùng dữ liệu. Xử lý bằng rule-based
47	FIELD_51	numerical	32.36%	Encode từ data liên quan đến thu nhập hay tiền đóng bảo hiểm? FIELD_50 đến 53 có thể liên quan với nhau theo thời gian	- Giá trị phổ biến nhất (4.413) là có vấn đề. Khi gộp cả train và test thì phát hiện ra giá trị này xuất hiện đúng 4412 lần (tức là bằng 4.413x1000-1) - Các giá trị còn lại đều nằm trong khoảng 60~67	
48	FIELD_52	numerical	32.36%	Encode từ data liên quan đến thu nhập hay tiền đóng bảo hiểm? FIELD_50 đến 53 có thể liên quan với nhau theo thời gian	- Giá trị phổ biến nhất (30.955) là có vấn đề. Khi gộp cả train và test thì phát hiện ra giá trị này xuất hiện đúng 30955 lần (tức là bằng 30.955x1000-1) - Các giá trị còn lại đều nằm trong khoảng 60~67	
49	FIELD_53	numerical	32.36%	Encode từ data liên quan đến thu nhập hay tiền đóng bảo hiểm? FIELD_50 đến 53 có thể liên quan với nhau theo thời gian	- Giá trị phổ biến nhất (31.171) là có vấn đề. Khi gộp cả train và test thì phát hiện ra giá trị này xuất hiện đúng 31170 lần (tức là bằng 31.171x1000-1) - Các giá trị còn lại đều nằm trong khoảng 60~67	
50	FIELD_54	numerical	32.36%		Encode của FIELD_50: nan -> nan, 29.77-> 0.0, 60.946 -> 0.12, 63.293 -> 0.25, 65.068 -> 00.38, 66.366 -> 0.5, 67.097 -> 0.62, 67.477 -> 0.75, 67.610 -> 0.88, 67.634 -> 1.	

#	Trường dữ liệu	Kiểu dữ liệu	Tỷ lệ missing	Ý nghĩa trường (Dự đoán)	Nhận xét (Lỗi, phân bố)	Phương án xử lý & chuẩn hoá (Cách thức chuẩn hoá, sửa lỗi)
51	FIELD_55	numerical	32.36%		Encode của FIELD_51 (tương tự trên)	
52	FIELD_56	numerical	32.36%		Encode của FIELD_52 (tương tự trên)	
53	FIELD_57	numerical	32.36%		Encode của FIELD_53 (tương tự trên)	
54	FIELD_6	numerical	1.58%		Giá trị discrete từ 0 đến 6 và nan (ko có giá trị 5). Giá trị 6 có trong test nhưng không có trong train. Phân bổ tương đối giống nhau giữa train và test. Giá trị 0 chiếm chủ yếu (88%) ở train và test.	
55	FIELD_7	categorical	0.00%	Lịch sử đóng bảo hiểm y tế (3340/BHXH-ST ngày 08/8/2017 của Bảo hiểm xã hội việt nam)	Có 37 mã unique trong train, 42 mã unique trong test, tổng 44 mã unique cả 2 tập. Mỗi record có thể nhận 0 hoặc nhiều giá trị, mỗi giá trị có thể lặp lại. Record có dãy giá trị dài nhất gồm 14 giá trị trong train và 14 giá trị trong test. Phân bố các giá trị tương đối giống nhau trong train và test, với các giá trị phổ biến nhất DN, GD, TE, HS, HC chiếm tộng cộng ~ 80%. Các mã hay xuất hiện cùng nhau nhất: GD-HS, GD-TE, DN-TE, DN-HS (trước và sau khi normalized)	1. Embedding cả string 2. Biến đổi string về array rồi embedding với sequence length = 14 3. Target encoding 4. Convert về bag of word count (multi-hot) 5. Convert về one-hot 6. Replace chuỗi thành 1 giá trị duy nhất (ví dụ thành giá trị phổ biến nhất hoặc ít phổ biến nhất 7. Rule-based
56	FIELD_8	categorical	32.36%	Giới tính: Nam, nữ	Gồm FEMALE, MALE và NaN. Phân bổ các giá trị (NaN, Female, Male) trên Train và Test lần lượt là (9678, 9291, 11031) và (6504, 6086, 7410)	Xử lý missing và để dạng categorical
57	FIELD_9	categorical	37,6%	Mã bảo hiểm y tế trong năm	Gần giống Field 7, tuy nhiên có xuất hiện thêm giá trị số '74', '75', '79', '80', '86'	
58	age_source1, age_source2	numerical	43.052%, 32.365%	Tuổi người tham gia bảo hiểm y tế và tuổi trong hồ sơ cho vay	2 trường age_source1, age_source2 là dữ liệu tuổi, hầu hết có dạng: + 2 tuổi bằng nhau + 1 trong 2 trường dữ liệu có tuổi + cả 2 không có tuổi + cả 2 có tuổi nhưng khác nhau Vấn đề: Một số tuổi trong tập test không có trong tập train	- Tạo một feature age_group để phân nhóm tuổi để bao quát được tất cả các tuổi mà có trong tập test . Việc phân nhóm tuổi theo từng giai đoạn của một người: 15-17: học sinh cấp 3, 18-21: sinh viên hoặc bắt đầu đi làm, 22-24: tốt nghiệp và bắt đầu đi làm, 25-27: ổn định công việc, có thể có gia đình, 28-30: có gia đình và lo con cái ..., 31-34, 35-37, 38 -40, 41-43, 44-46, 47-49, 50-53, 53-55, 55-57, 57-59, 60-65, 65-69, 70-79, 80-99 ko xd: dữ liệu missing
59	district	categorical	43.08%	Quận, huyện theo tỉnh, thành phố	- Các huyện tên chưa đồng nhất vì khác nhau chữ hoa, chữ thường - Một số huyện trùng tên ở các tỉnh khác nhau, như huyện Châu Thành có ở 9 tỉnh ở miền nam (Long An, Tiền Giang, Bến Tre, An Giang, Đồng Tháp, Hậu Giang (2 huyện Châu Thành và Châu Thành A), Kiên Giang, Trà Vinh và Tây Ninh.)	- Chuẩn hóa lại dữ liệu thành chữ thường - Xử lý trường quận huyện đi kèm với tên tỉnh district = district + province - Sử dụng pre-trained embeddings

#	Trường dữ liệu	Kiểu dữ liệu	Tỷ lệ missing	Ý nghĩa trường (Dự đoán)	Nhận xét (Lỗi, phân bố)	Phương án xử lý & chuẩn hoá (Cách thức chuẩn hoá, sửa lỗi)
60	maCv	categorical	32.37%	Công việc	- Dữ liệu None lớn (hơn 13.000). missing lớn (hơn 9.600) - Khoảng hơn 7000 mã công việc nhưng chủ yếu là công nhân, giáo viên, nhân viên. - Chú yếu bad credit ở công nhân - Dữ liệu viết tắt và sai chính tả	- Phân nhóm công việc - Xác định một số nhóm good credit như bác sỹ, trưởng công an, hiệu trưởng,... - Sử dụng pre-trained embeddings
61	province	categorical	43.08%	Tỉnh, thành phố	Thực tế Việt Nam có 63 tỉnh, thành phố trên cả nước. Trong dữ liệu có 64 tên tỉnh thành phố do tên Tỉnh Hòa Bình và Tình Hoà Bình bị sai chính tả nên thành 2 tỉnh.	Chuyển dữ liệu thành chữ thường và chuẩn hóa lại lỗi chính tả để thành 63 tỉnh, thành phố - Sử dụng pre-train embeddings