**XÂY DỰNG SCORECARD VỚI DỮ LIỆU CỦA LENDING CLUB**

**I. Tổng quan dữ liệu**

Bài nghiên cứu ứng dụng mô hình Logistic chấm điểm 115.679 khách hàng cá nhân nộp hồ sơ vay vốn tại Lending Club - tổ chức cho vay ngang hàng tại Mỹ. Bộ dữ liệu của Lending Club cung cấp 144 biến đặc điểm của khách hàng vay vốn bao gồm các loại biến sau:

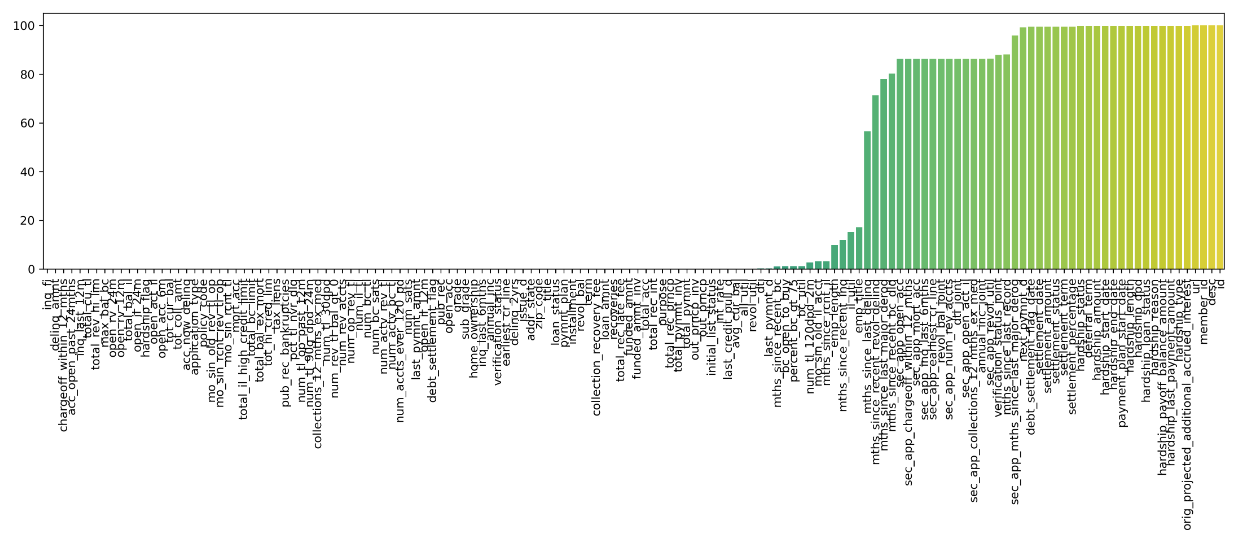
|  |  |
| --- | --- |
| **Thông tin cá nhân** | Tình trạng việc làm, nghề nghiệp, thu nhập, nơi sinh sống, số lượng người phụ thuộc… |
| **Lịch sử tín dụng gần đây** | Chiều dài của lịch sử tín dụng, số lượng và giá trị của quá khứ vay vốn, số lượng và giá trị của các khoản vay xấu trong quá khứ… |
| **Hành vi** | Lịch sử về số tiền chi tiêu, việc trả nợ thực tế… |

**II. Lựa chọn biến ban đầu**

**1. Phân tích biến**

**1.1. Tỷ lệ dữ liệu thiếu**

**Biểu đồ 1: Biểu đồ tỷ lệ dữ liệu thiếu**



Dựa trên kết quả kiểm tra dữ liệu, 43 biến bị loại khỏi dữ liệu xây dựng mô hình do tỷ lệ giá trị còn thiếu cao (>70% giá trị bị thiếu).

**Bảng 1: 43 biến bị loại vì tỷ lệ dữ liệu thiếu cao**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Biến** | **Ý nghĩa** | **% Thiếu dữ liệu** |
| 1 | id | ID của khoản vay | 100.00 |
| 2 | desc | Mô tả khoản vay do người đi vay cung cấp | 100.00 |
| 3 | member\_id | ID của người vay | 100.00 |
| 4 | url | Đường dẫn của khoản vay trên trang LC | 100.00 |
| 5 | orig\_projected\_additional\_accrued\_interest | NaN | 99.88 |
| 6 | hardship\_type | NaN | 99.87 |
| 7 | hardship\_last\_payment\_amount | NaN | 99.87 |
| 8 | hardship\_payoff\_balance\_amount | NaN | 99.87 |
| 9 | hardship\_reason | NaN | 99.87 |
| 10 | hardship\_loan\_status | NaN | 99.87 |
| 11 | hardship\_dpd | NaN | 99.87 |
| 12 | hardship\_length | NaN | 99.87 |
| 13 | payment\_plan\_start\_date | NaN | 99.87 |
| 14 | hardship\_end\_date | NaN | 99.87 |
| 15 | hardship\_start\_date | NaN | 99.87 |
| 16 | hardship\_amount | NaN | 99.87 |
| 17 | deferral\_term | NaN | 99.87 |
| 18 | hardship\_status | NaN | 99.87 |
| 19 | settlement\_percentage | NaN | 99.48 |
| 20 | settlement\_term | NaN | 99.48 |
| 21 | settlement\_status | NaN | 99.48 |
| 22 | settlement\_amount | NaN | 99.48 |
| 23 | settlement\_date | NaN | 99.48 |
| 24 | debt\_settlement\_flag\_date | NaN | 99.48 |
| 25 | next\_pymnt\_d | Ngày trả nợ tiếp theo | 99.25 |
| 26 | sec\_app\_mths\_since\_last\_major\_derog | NaN | 95.87 |
| 27 | mths\_since\_last\_record | Số tháng kể từ bản ghi công khai cuối cùng. | 88.10 |
| 28 | verification\_status\_joint | Indicates if income was verified by LC, not verified, or if the income source was verified | 87.85 |
| 29 | sec\_app\_revol\_util | NaN | 86.40 |
| 30 | annual\_inc\_joint | Thu nhập hàng năm được KH báo cáo kết hợp với những người đồng vay cung cấp trong quá trình đăng ký | 86.21 |
| 31 | sec\_app\_collections\_12\_mths\_ex\_med | NaN | 86.21 |
| 32 | sec\_app\_open\_act\_il | NaN | 86.21 |
| 33 | dti\_joint | Hệ số nợ trên thu nhập | 86.21 |
| 34 | sec\_app\_num\_rev\_accts | NaN | 86.21 |
| 35 | revol\_bal\_joint | NaN | 86.21 |
| 36 | sec\_app\_earliest\_cr\_line | NaN | 86.21 |
| 37 | sec\_app\_inq\_last\_6mths | NaN | 86.21 |
| 38 | sec\_app\_mort\_acc | NaN | 86.21 |
| 39 | sec\_app\_chargeoff\_within\_12\_mths | NaN | 86.21 |
| 40 | sec\_app\_open\_acc | NaN | 86.21 |
| 41 | mths\_since\_recent\_bc\_dlq | Các tháng kể từ lần để nợ quá hạn thẻ ngân hàng gần nhất | 80.24 |
| 42 | mths\_since\_last\_major\_derog | Các tháng kể từ lần bị xếp hạng nợ quá hạn 90 ngày hoặc tệ hơn | 78.02 |
| 43 | mths\_since\_recent\_revol\_delinq | Các tháng kể từ lần dể nợ quá hạn sản phẩm revolving | 71.27 |

\*NaN: không có trong từ điển dữ liệu

**1.2. Mức độ có liên quan**

26 biến bị loại khỏi mục đích xây dựng mô hình vì chúng cung cấp thông tin quá cụ thể về người vay hoặc khoản vay nên không thể áp dụng chung cho mọi trường hợp, hoặc không liên quan đến mục đích xây dựng mô hình, hoặc đã có biến khác tương tự.

**Bảng 2: 26 biến bị loại vì không liên quan**

|  |  |  |
| --- | --- | --- |
| **#** | **Biến** | **Ý nghĩa** |
| 1 | sub\_grade | Mức điểm chi tiết hơn do LC chấm |
| 2 | emp\_title | Chức danh công việc |
| 3 | pymnt\_plan | Cho biết kế hoạch thanh toán đã được đưa ra cho khoản vay hay chưa |
| 4 | title | Tiêu đề khoản vay do người đi vay cung cấp |
| 5 | zip\_code | Mã vùng |
| 6 | total\_rec\_late\_fee | Phí quá hạn nhận được cho đến nay |
| 7 | recoveries | Tổng thu hồi sau khoanh nợ |
| 8 | collection\_recovery\_fee | Phí thu hồi nợ sau khoanh nợ |
| 9 | last\_pymnt\_d | Tháng mà lần cuối thanh toán |
| 10 | last\_pymnt\_amnt | Số tiền thanh toán lần gần nhất |
| 11 | last\_credit\_pull\_d | Tháng gần đây nhất rút tín dụng cho khoản vay |
| 12 | collections\_12\_mths\_ex\_med | Số lần thu hồi trong 12 tháng không bao gồm y tế |
| 13 | policy\_code | Mã chính sách |
| 14 | acc\_now\_delinq | Số tài khoản mà người vay hiện đang quá hạn |
| 15 | num\_tl\_120dpd\_2m | Số tài khoản hiện tại đã quá hạn 120 ngày |
| 16 | num\_tl\_30dpd | Số tài khoản hiện tại đã quá hạn 30 ngày |
| 17 | hardship\_flag | NaN |
| 18 | funded\_amnt | Tổng số tiền cam kết cho khoản vay đó tại thời điểm đó |
| 19 | funded\_amnt\_inv | Tổng số tiền nhà đầu tư cam kết cho khoản vay đó tại thời điểm đó |
| 20 | out\_prncp\_inv | Nợ gốc còn lại cho một phần của tổng số tiền được tài trợ bởi các nhà đầu tư |
| 21 | total\_pymnt | Các khoản thanh toán nhận được cho đến nay |
| 22 | total\_pymnt\_inv | Các khoản thanh toán nhận được cho đến nay cho một phần của tổng số tiền được tài trợ bởi các nhà đầu tư |
| 23 | out\_prncp | Số tiền gốc còn lại cho tổng số tiền đã tài trợ |
| 24 | total\_rec\_prncp | Số tiền gốc nhận được cho đến nay |
| 25 | total\_rec\_int | Tiền lãi nhận được cho đến nay |
| 26 | mths\_since\_last\_delinq | Số tháng kể từ lần vi phạm cuối cùng của người vay |

**1.3. Mức độ tập trung**

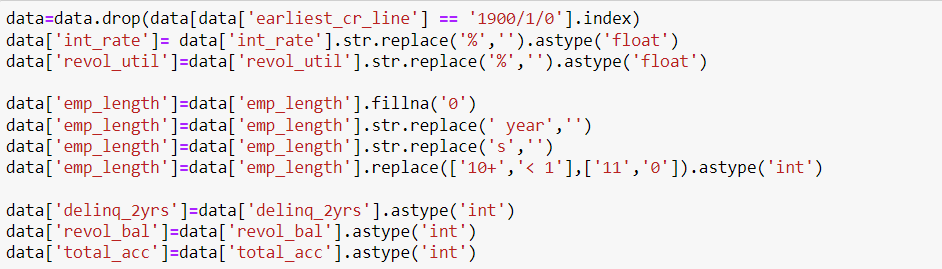
Dựa trên kết quả kiểm tra dữ liệu, 2 biến sau bị loại trong dữ liệu xây dựng mô hình vì hầu hết các quan sát đều có cùng giá trị.

**Bảng 3: 2 biến bị loại do mức độ tập trung cao**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Biến** | **Ý nghĩa** | **Mức độ tập trung** |
| 1 | debt\_settlement\_flag | Có đảo nợ hay không | 99.5% - N |
| 2 | tax\_liens | Số quyền nắm giữ tài sản vì nợ thuế | 100% - 0 |

**2. Làm sạch dữ liệu**

Loại bỏ các ký tự đặc biệt trong dữ liệu (%, +, -, <, >) hoặc các từ đi kèm (‘year’, ‘s’) và chuyển đổi dữ liệu về dạng ‘float’ hoặc ‘int’ ở các biến cần sử dụng dữ liệu dạng số. Loại bỏ các năm quá xa (1900). Gán định dạng datetime cho các biến dạng ngày tháng.



**3. Tạo các biến mới**

Một số biến có sẵn trong dữ liệu không phù hợp để xây dựng mô hình, 2 biến bổ sung được tạo ra từ những dữ liệu này để thêm góc nhìn khác về khách hàng.

**Bảng 4: 2 biến mới được tạo ra**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Biến tạo mới** | **Ý nghĩa** | **Logic tạo biến** |
| 1 | mth\_interval | Số tháng kể từ lần đi vay lần đầu đến lần đi vay hiện tại | data['issue\_d']-data['earliest\_cr\_line'] |
| 2 | income\_vs\_loan | Tỉ lệ thu nhập trên khoản nợ | data['annual\_inc']/data['loan\_amnt'] |

Sau khi tạo ra các biến tạo mới, 2 biến issue\_d và earliest\_cr\_line bị bỏ trong dữ liệu xây dựng mô hình

**4. Trước khi xây dựng mô hình**

**4.1. Khả năng dự báo (IV)**

Chỉ những biến có khả năng dự báo đáng kể (IV > 0.02) sẽ được xem xét để phát triển mô hình. Một số biến có IV < 0.02 nhưng > 0.01 và có ý nghĩa quan trọng sẽ được giữ lại (như annual\_inc và addr\_state)

Dựa trên giá trị IV được quan sát, các biến dưới đây sẽ được loại trừ khỏi dữ liệu phát triển mô hình do khả năng dự báo thấp.

**Bảng 5: Các biến bị loại do IV thấp**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Biến** | **Ý nghĩa** | **IV** |
| 1 | purpose |  | 0.0093 |
| 2 | dti |  | 0.0110 |
| 3 | delinq\_2yrs |  | 0.0002 |
| 4 | open\_acc |  | 0.0016 |
| 5 | pub\_rec |  | 0.0007 |
| 6 | revol\_bal |  | 0.0059 |
| 7 | revol\_util |  | 0.0039 |
| 8 | total\_acc |  | 0.0024 |
| 9 | initial\_list\_status |  | 0.0010 |
| 10 | application\_type |  | 0.0002 |
| 11 | tot\_coll\_amt |  | 0.0007 |
| 12 | open\_act\_il |  | 0.0065 |
| 13 | open\_il\_12m |  | 0.0052 |
| 14 | open\_il\_24m |  | 0.0045 |
| 15 | mths\_since\_rcnt\_il |  | 0.0063 |
| 16 | total\_bal\_il |  | 0.0025 |
| 17 | il\_util |  | 0.0102 |
| 18 | max\_bal\_bc |  | 0.0131 |
| 19 | all\_util |  | 0.0107 |
| 20 | total\_rev\_hi\_lim |  | 0.0148 |
| 21 | inq\_fi |  | 0.0125 |
| 22 | total\_cu\_tl |  | 0.0008 |
| 23 | bc\_open\_to\_buy |  | 0.0173 |
| 24 | bc\_util |  | 0.0040 |
| 25 | chargeoff\_within\_12\_mths |  | 0.0001 |
| 26 | delinq\_amnt |  | 0.0001 |
| 27 | mo\_sin\_old\_il\_acct |  | 0.0160 |
| 28 | mo\_sin\_rcnt\_tl |  | 0.0162 |
| 29 | num\_accts\_ever\_120\_pd |  | 0.0011 |
| 30 | num\_actv\_bc\_tl |  | 0.0082 |
| 31 | num\_actv\_rev\_tl |  | 0.0142 |
| 32 | num\_bc\_sats |  | 0.0054 |
| 33 | num\_bc\_tl |  | 0.0016 |
| 34 | num\_il\_tl |  | 0.0096 |
| 35 | num\_op\_rev\_tl |  | 0.0081 |
| 36 | num\_rev\_accts |  | 0.0020 |
| 37 | num\_rev\_tl\_bal\_gt\_0 |  | 0.0123 |
| 38 | num\_sats |  | 0.0017 |
| 39 | num\_tl\_90g\_dpd\_24m |  | 0.0004 |
| 40 | pct\_tl\_nvr\_dlq |  | 0.0013 |
| 41 | percent\_bc\_gt\_75 |  | 0.0064 |
| 42 | pub\_rec\_bankruptcies |  | 0.0007 |
| 43 | total\_bal\_ex\_mort |  | 0.0042 |
| 44 | total\_bc\_limit |  | 0.0185 |
| 45 | total\_il\_high\_credit\_limit |  | 0.0053 |
| 46 | mth\_interval |  | 0.0089 |
| 47 | issue\_m |  | 0.0003 |

**4.2. Binning**

Bảng: Kết quả Binning

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Biến** | **Bin** | **Số lượng** | **%Số lượng** | **Số lượng Good** | **Số lượng Bad** | **%Bad** | **WOE** |
| loan\_amnt | [-inf,10000.0) | 30087 | 0.2963 | 27070 | 3017 | 0.1003 | -0.3803 |
| [10000.0,19000.0) | 36817 | 0.3626 | 31549 | 5268 | 0.1431 | 0.0239 |
| [19000.0,inf) | 34633 | 0.3411 | 28685 | 5948 | 0.1717 | 0.2405 |
| term | 36 months | 77498 | 0.7632 | 68989 | 8509 | 0.1098 | -0.2790 |
| 60 months | 24039 | 0.2368 | 18315 | 5724 | 0.2381 | 0.6508 |
| int\_rate | [-inf,9.0) | 34886 | 0.3436 | 32715 | 2171 | 0.0622 | -0.8988 |
| [9.0,12.0) | 20060 | 0.1976 | 17775 | 2285 | 0.1139 | -0.2376 |
| [12.0,17.0) | 29337 | 0.2889 | 24049 | 5288 | 0.1803 | 0.2992 |
| [17.0,inf) | 17254 | 0.1699 | 12765 | 4489 | 0.2602 | 0.7688 |
| installment | [-inf,200.0) | 17062 | 0.1680 | 15513 | 1549 | 0.0908 | -0.4902 |
| [200.0,500.0) | 46645 | 0.4594 | 40182 | 6463 | 0.1386 | -0.0135 |
| [500.0,650.0) | 13241 | 0.1304 | 11145 | 2096 | 0.1583 | 0.1429 |
| [650.0,1150.0) | 20355 | 0.2005 | 16779 | 3576 | 0.1757 | 0.2679 |
| [1150.0,inf) | 4234 | 0.0417 | 3685 | 549 | 0.1297 | -0.0901 |
| grade | A | 34864 | 0.3434 | 32694 | 2170 | 0.0622 | -0.8986 |
| B | 27937 | 0.2751 | 24435 | 3502 | 0.1254 | -0.1288 |
| C | 23078 | 0.2273 | 18661 | 4417 | 0.1914 | 0.3729 |
| D%,%E%,%F%,%G | 15658 | 0.1542 | 11514 | 4144 | 0.2647 | 0.7919 |
| emp\_length | [-inf,1.0) | 23289 | 0.2294 | 19198 | 4091 | 0.1757 | 0.2678 |
| [1.0,4.0) | 23316 | 0.2296 | 20069 | 3247 | 0.1393 | -0.0076 |
| [4.0,inf) | 54932 | 0.5410 | 48037 | 6895 | 0.1255 | -0.1273 |
| home\_ownership | NONE, MORTGAGE | 50811 | 0.5004 | 44827 | 5984 | 0.1178 | -0.1999 |
| OWN | 11718 | 0.1154 | 9928 | 1790 | 0.1528 | 0.1007 |
| RENT,ANY | 39008 | 0.3842 | 32549 | 6459 | 0.1656 | 0.1966 |
| annual\_inc | [-inf,40000.0) | 15083 | 0.1485 | 12615 | 2468 | 0.1636 | 0.1824 |
| [40000.0,100000.0) | 59859 | 0.5895 | 51320 | 8539 | 0.1427 | 0.0204 |
| [100000.0,170000.0) | 20151 | 0.1985 | 17672 | 2479 | 0.1230 | -0.1503 |
| [170000.0,inf) | 6444 | 0.0635 | 5697 | 747 | 0.1159 | -0.2178 |
| verification\_status | Not Verified | 50224 | 0.4946 | 43905 | 6319 | 0.1258 | -0.1246 |
| Source Verified | 35534 | 0.3500 | 30466 | 5068 | 0.1426 | 0.0202 |
| Verified | 15779 | 0.1554 | 12933 | 2846 | 0.1804 | 0.3000 |
| addr\_state | ME,WV,WY,DE,DC,NH,OR,ID,MI,AK,VT,  CO,WA,MT,SC,RI,VA,WI | 17465 | 0.1720 | 15414 | 2051 | 0.1174 | -0.2031 |
| OH,MN,NC,MO,KY,UT,IL,IN,GA | 19932 | 0.1963 | 17373 | 2559 | 0.1284 | -0.1015 |
| MA,NJ,AZ,AR,TX,MD | 19840 | 0.1954 | 17062 | 2778 | 0.1400 | -0.0013 |
| AL,CA,CT,NM,TN | 19183 | 0.1889 | 16396 | 2787 | 0.1453 | 0.0418 |
| PA,LA,KS,NY,MS,OK,FL,NE,SD,HI,ND,NV | 25117 | 0.2474 | 21059 | 4058 | 0.1616 | 0.1672 |
| inq\_last\_6mths | [-inf,1.0) | 67593 | 0.6657 | 58957 | 8636 | 0.1278 | -0.1070 |
| [1.0,2.0) | 24763 | 0.2439 | 20927 | 3836 | 0.1549 | 0.1172 |
| [2.0,inf) | 9181 | 0.0904 | 7420 | 1761 | 0.1918 | 0.3755 |
| tot\_cur\_bal | [-inf,100000.0) | 54774 | 0.5394 | 46023 | 8751 | 0.1598 | 0.1539 |
| [100000.0,460000.0) | 41156 | 0.4053 | 36242 | 4914 | 0.1194 | -0.1843 |
| [460000.0,inf) | 5607 | 0.0552 | 5039 | 568 | 0.1013 | -0.3690 |
| open\_acc\_6m | [-inf,1.0) | 46061 | 0.4536 | 40302 | 5759 | 0.1250 | -0.1318 |
| [1.0,2.0) | 31422 | 0.3095 | 27003 | 4419 | 0.1406 | 0.0038 |
| [2.0,3.0) | 14686 | 0.1446 | 12316 | 2370 | 0.1614 | 0.1658 |
| [3.0,inf) | 9368 | 0.0923 | 7683 | 1685 | 0.1799 | 0.2966 |
| open\_rv\_12m | [-inf,1.0) | 40855 | 0.4024 | 35819 | 5036 | 0.1233 | -0.1480 |
| [1.0,2.0) | 29711 | 0.2926 | 25623 | 4088 | 0.1376 | -0.0216 |
| [2.0,3.0) | 15980 | 0.1574 | 13576 | 2404 | 0.1504 | 0.0827 |
| [3.0,4.0) | 7839 | 0.0772 | 6515 | 1324 | 0.1689 | 0.2204 |
| [4.0,inf) | 7152 | 0.0704 | 5771 | 1381 | 0.1931 | 0.3838 |
| open\_rv\_24m | [-inf,2.0) | 42589 | 0.4194 | 37531 | 5058 | 0.1188 | -0.1904 |
| [2.0,3.0) | 19274 | 0.1898 | 16698 | 2576 | 0.1337 | -0.0552 |
| [3.0,5.0) | 23011 | 0.2266 | 19556 | 3455 | 0.1501 | 0.0804 |
| [5.0,inf) | 16663 | 0.1641 | 13519 | 3144 | 0.1887 | 0.3552 |
| inq\_last\_12m | [-inf,1.0) | 31507 | 0.3103 | 27763 | 3744 | 0.1188 | -0.1897 |
| [1.0,2.0) | 25419 | 0.2503 | 22086 | 3333 | 0.1311 | -0.0772 |
| [2.0,4.0) | 27139 | 0.2673 | 23056 | 4083 | 0.1504 | 0.0827 |
| [4.0,inf) | 17472 | 0.1721 | 14399 | 3073 | 0.1759 | 0.2693 |
| acc\_open\_past\_24mths | [-inf,3.0) | 31498 | 0.3102 | 27837 | 3661 | 0.1162 | -0.2148 |
| [3.0,4.0) | 15464 | 0.1523 | 13499 | 1965 | 0.1271 | -0.1133 |
| [4.0,8.0) | 39933 | 0.3933 | 34026 | 5907 | 0.1479 | 0.0628 |
| [8.0,inf) | 14642 | 0.1442 | 11942 | 2700 | 0.1844 | 0.3270 |
| avg\_cur\_bal | [-inf,8000.0) | 52396 | 0.5160 | 43948 | 8448 | 0.1612 | 0.1648 |
| [8000.0,16000.0) | 17735 | 0.1747 | 15404 | 2331 | 0.1314 | -0.0745 |
| [16000.0,36000.0) | 22056 | 0.2172 | 19522 | 2534 | 0.1149 | -0.2279 |
| [36000.0,inf) | 9350 | 0.0921 | 8430 | 920 | 0.0984 | -0.4013 |
| mo\_sin\_old\_rev\_tl\_op | [-inf,60.0) | 9841 | 0.0969 | 8051 | 1790 | 0.1819 | 0.3103 |
| [60.0,80.0) | 7323 | 0.0721 | 6094 | 1229 | 0.1678 | 0.2127 |
| [80.0,140.0) | 23717 | 0.2336 | 20135 | 3582 | 0.1510 | 0.0873 |
| [140.0,170.0) | 16108 | 0.1586 | 13905 | 2203 | 0.1368 | -0.0286 |
| [170.0,inf) | 44548 | 0.4387 | 39119 | 5429 | 0.1219 | -0.1610 |
| mo\_sin\_rcnt\_rev\_tl\_op | [-inf,7.0) | 40034 | 0.3943 | 33735 | 6299 | 0.1573 | 0.1357 |
| [7.0,17.0) | 30297 | 0.2984 | 26056 | 4241 | 0.1400 | -0.0016 |
| [17.0,35.0) | 20644 | 0.2033 | 18081 | 2563 | 0.1242 | -0.1399 |
| [35.0,inf) | 10562 | 0.1040 | 9432 | 1130 | 0.1070 | -0.3081 |
| mort\_acc | [-inf,1.0) | 45724 | 0.4503 | 38309 | 7415 | 0.1622 | 0.1717 |
| [1.0,2.0) | 18252 | 0.1798 | 15774 | 2478 | 0.1358 | -0.0371 |
| [2.0,3.0) | 15599 | 0.1536 | 13690 | 1909 | 0.1224 | -0.1563 |
| [3.0,inf) | 21962 | 0.2163 | 19531 | 2431 | 0.1107 | -0.2699 |
| mths\_since\_recent\_bc | [-inf,6.0) | 21983 | 0.2165 | 18416 | 3567 | 0.1623 | 0.1723 |
| [6.0,18.0) | 33243 | 0.3274 | 28330 | 4913 | 0.1478 | 0.0618 |
| [18.0,30.0) | 18808 | 0.1852 | 16256 | 2552 | 0.1357 | -0.0378 |
| [30.0,50.0) | 14388 | 0.1417 | 12568 | 1820 | 0.1265 | -0.1185 |
| [50.0,inf) | 13115 | 0.1292 | 11734 | 1381 | 0.1053 | -0.3258 |
| mths\_since\_recent\_inq | [-inf,3.0) | 21355 | 0.2103 | 17630 | 3725 | 0.1744 | 0.2593 |
| [3.0,5.0) | 13262 | 0.1306 | 11249 | 2013 | 0.1518 | 0.0932 |
| [5.0,7.0) | 11840 | 0.1166 | 10142 | 1698 | 0.1434 | 0.0266 |
| [7.0,inf) | 55080 | 0.5425 | 48283 | 6797 | 0.1234 | -0.1468 |
| num\_tl\_op\_past\_12m | [-inf,2.0) | 47615 | 0.4689 | 41767 | 5848 | 0.1228 | -0.1522 |
| [2.0,3.0) | 21488 | 0.2116 | 18484 | 3004 | 0.1398 | -0.0031 |
| [3.0,4.0) | 14073 | 0.1386 | 11898 | 2175 | 0.1546 | 0.1145 |
| [4.0,5.0) | 8374 | 0.0825 | 7016 | 1358 | 0.1622 | 0.1717 |
| [5.0,inf) | 9987 | 0.0984 | 8139 | 1848 | 0.1850 | 0.3313 |
| tot\_hi\_cred\_lim | [-inf,160000.0) | 57291 | 0.5642 | 48153 | 9138 | 0.1595 | 0.1519 |
| [160000.0,460000.0) | 35186 | 0.3465 | 30992 | 4194 | 0.1192 | -0.1862 |
| [460000.0,inf) | 9060 | 0.0892 | 8159 | 901 | 0.0994 | -0.3895 |
| income\_vs\_loan | [-inf,2.5) | 11608 | 0.1143 | 9177 | 2431 | 0.2094 | 0.4854 |
| [2.5,4.5) | 30351 | 0.2989 | 25292 | 5059 | 0.1667 | 0.2045 |
| [4.5,8.0) | 31015 | 0.3055 | 27030 | 3985 | 0.1285 | -0.1006 |
| [8.0,inf) | 28563 | 0.2813 | 25805 | 2758 | 0.0966 | -0.4222 |

Các biến bị loại bỏ do không có tính logic, hợp lý về mặt vận hành, ứng dụng:

|  |  |  |
| --- | --- | --- |
| **#** | **Biến** | **Ý nghĩa** |
| 1 | installment |  |
| 2 | num\_tl\_90g\_dpd\_24m |  |
| 3 | chargeoff\_within\_12mths |  |
| 4 | num\_sats |  |
| 5 | pct\_tl\_nvr\_dlq |  |
| 6 | open\_acc |  |
| 7 | delinq\_amnt |  |
| 8 | total\_bal\_il |  |
| 9 | percent\_bc\_gt\_75 |  |
| 10 | revol\_util |  |
| 11 | mths\_since\_rcnt\_il |  |

**4.2. Tương quan**

Các cặp biến có mức độ tương quan mạnh (từ 0.6 - 1):

|  |  |  |
| --- | --- | --- |
| **Biến 1** | **Biến 2** | **Mức độ tương quan** |
| mths\_since\_recent\_bc | mo\_sin\_rcnt\_rev\_tl\_op | 0.62 |
| open\_acc\_6m | open\_rv\_12m | 0.64 |
| open\_rv\_12m | acc\_open\_past\_24mths | 0.67 |
| num\_tl\_op\_past\_12m | open\_rv\_24m | 0.67 |
| num\_tl\_op\_past\_12m | open\_acc\_6m | 0.75 |
| tot\_hi\_cred\_lim | avg\_cur\_bal | 0.76 |
| open\_rv\_12m | open\_rv\_24m | 0.79 |
| acc\_open\_past\_24mths | num\_tl\_op\_past\_12m | 0.79 |
| avg\_cur\_bal | tot\_cur\_bal | 0.81 |
| num\_tl\_op\_past\_12m | open\_rv\_12m | 0.83 |
| acc\_open\_past\_24mths | open\_rv\_24m | 0.84 |
| tot\_hi\_cred\_lim | tot\_cur\_bal | 0.97 |

Cặp biến cuối cùng có hệ số tương quan vượt ngưỡng cho phép (0.85). Trong 2 biến này, biến tot\_hi\_cred\_lim là biến có IV cao hơn, nên loại bỏ biến tot\_cur\_bal khỏi dữ liệu xây dựng mô hình.

**5. Xây dựng mô hình**

**5.1. Hồi quy Logistic**

Hồi quy cho ra bảng kết quả như sau:

|  |  |
| --- | --- |
| **Biến** | **Hệ số** |
| **Hệ số chặn** | **-1.8267** |
| emp\_length\_woe | 1.1045 |
| int\_rate\_woe | 0.2893 |
| mort\_acc\_woe | 0.0881 |
| inq\_last\_12m\_woe | 0.3252 |
| open\_rv\_24m\_woe | 0.4524 |
| term\_woe | 0.6038 |
| verification\_status\_woe | 0.2288 |
| loan\_amnt\_woe | 0.6992 |
| open\_acc\_6m\_woe | 0.2385 |
| avg\_cur\_bal\_woe | 0.1740 |
| grade\_woe | 0.4815 |
| num\_tl\_op\_past\_12m\_woe | 0.0000 |
| home\_ownership\_woe | 0.6758 |
| inq\_last\_6mths\_woe | 0.1971 |
| open\_rv\_12m\_woe | -0.1328 |
| acc\_open\_past\_24mths\_woe | 0.2594 |
| tot\_hi\_cred\_lim\_woe | 0.4593 |
| addr\_state\_woe | 1.0054 |
| mo\_sin\_rcnt\_rev\_tl\_op\_woe | -0.0614 |
| mths\_since\_recent\_bc\_woe | 0.3388 |
| mths\_since\_recent\_inq\_woe | 0.3617 |
| mo\_sin\_old\_rev\_tl\_op\_woe | 0.4800 |
| income\_vs\_loan\_woe | 0.2828 |
| annual\_inc\_woe | 0.2578 |

**5.2. Tính VIF**

|  |  |
| --- | --- |
| **Biến** | **VIF** |
| emp\_length\_woe | 1.050 |
| int\_rate\_woe | 20.412 |
| mort\_acc\_woe | 2.495 |
| inq\_last\_12m\_woe | 2.094 |
| open\_rv\_24m\_woe | 3.866 |
| term\_woe | 1.374 |
| verification\_status\_woe | 1.117 |
| loan\_amnt\_woe | 2.988 |
| open\_acc\_6m\_woe | 2.573 |
| avg\_cur\_bal\_woe | 3.519 |
| grade\_woe | 20.309 |
| num\_tl\_op\_past\_12m\_woe | 5.493 |
| home\_ownership\_woe | 2.026 |
| inq\_last\_6mths\_woe | 2.119 |
| open\_rv\_12m\_woe | 5.151 |
| acc\_open\_past\_24mths\_woe | 3.680 |
| tot\_hi\_cred\_lim\_woe | 3.725 |
| addr\_state\_woe | 1.019 |
| mo\_sin\_rcnt\_rev\_tl\_op\_woe | 2.746 |
| mths\_since\_recent\_bc\_woe | 1.952 |
| mths\_since\_recent\_inq\_woe | 2.330 |
| mo\_sin\_old\_rev\_tl\_op\_woe | 1.188 |
| income\_vs\_loan\_woe | 2.947 |

=> Loại bỏ các biến hệ số = 0 và VIF > 5, bao gồm các biến:

|  |
| --- |
| num\_tl\_op\_past\_12m |
| int\_rate |
| grade |
| open\_rv\_12m |

**5.2. Thực hiện hồi quy lại**

|  |  |
| --- | --- |
| **Biến** | **Hệ số** |
| emp\_length\_woe | 1.1244 |
| mort\_acc\_woe | 0.3114 |
| inq\_last\_12m\_woe | 0.4888 |
| open\_rv\_24m\_woe | 0.3336 |
| term\_woe | 0.9309 |
| verification\_status\_woe | 0.6421 |
| loan\_amnt\_woe | 0.5821 |
| open\_acc\_6m\_woe | 0.3011 |
| avg\_cur\_bal\_woe | -0.0101 |
| home\_ownership\_woe | 0.6687 |
| inq\_last\_6mths\_woe | 0.2860 |
| acc\_open\_past\_24mths\_woe | 0.4599 |
| tot\_hi\_cred\_lim\_woe | 0.7539 |
| addr\_state\_woe | 1.0369 |
| mo\_sin\_rcnt\_rev\_tl\_op\_woe | -0.1045 |
| mths\_since\_recent\_bc\_woe | 0.2748 |
| mths\_since\_recent\_inq\_woe | 0.4377 |
| mo\_sin\_old\_rev\_tl\_op\_woe | 0.6547 |
| income\_vs\_loan\_woe | 0.2763 |
| annual\_inc\_woe | 0.5231 |

**5.3. Thẻ điểm**

|  |  |  |
| --- | --- | --- |
| **Biến** | **Bin** | **Điểm** |
| basepoint |  | 415 |
| emp\_length | [-inf,1.0) | -22 |
| [1.0,4.0) | 1 |
| [4.0,inf) | 10 |
| mort\_acc | [-inf,1.0) | -4 |
| [1.0,2.0) | 1 |
| [2.0,3.0) | 4 |
| [3.0,inf) | 6 |
| inq\_last\_12m | [-inf,1.0) | 7 |
| [1.0,2.0) | 3 |
| [2.0,4.0) | -3 |
| [4.0,inf) | -9 |
| open\_rv\_24m | [-inf,2.0) | 5 |
| [2.0,3.0) | 1 |
| [3.0,5.0) | -2 |
| [5.0,inf) | -9 |
| term | 36 months | 19 |
| 60 months | -44 |
| verification\_status | Not Verified | 6 |
| Source Verified | -1 |
| Verified | -14 |
| loan\_amnt | [-inf,10000.0) | 16 |
| [10000.0,19000.0) | -1 |
| [19000.0,inf) | -10 |
| open\_acc\_6m | [-inf,1.0) | 3 |
| [1.0,2.0) | 0 |
| [2.0,3.0) | -4 |
| [3.0,inf) | -6 |
| home\_ownership | NONE, MORTGAGE | 10 |
| OWN | -5 |
| RENT, ANY | -9 |
| inq\_last\_6mths | [-inf,1.0) | 2 |
| inq\_last\_6mths | [1.0,2.0) | -2 |
| inq\_last\_6mths | [2.0,inf) | -8 |
| acc\_open\_past\_24mths | [-inf,3.0) | 7 |
| [3.0,4.0) | 4 |
| [4.0,8.0) | -2 |
| [8.0,inf) | -11 |
| tot\_hi\_cred\_lim | [-inf,160000.0) | -8 |
| [160000.0,460000.0) | 10 |
| [460000.0,inf) | 21 |
| addr\_state | ME,WV,WY,DE,DC,NH,OR,ID,MI,AK,VT,  CO,WA,MT,SC,RI,VA,WI | 15 |
| OH,MN,NC,MO,KY,UT,IL,IN,GA | 8 |
| MA,NJ,AZ,AR,TX,MD | 0 |
| AL,CA,CT,NM,TN | -3 |
| PA,LA,KS,NY,MS,OK,FL,NE,SD,HI,ND,NV | -13 |
| mo\_sin\_rcnt\_rev\_tl\_op | [-inf,7.0) | 1 |
| [7.0,17.0) | 0 |
| [17.0,35.0) | -1 |
| [35.0,inf) | -2 |
| mths\_since\_recent\_bc | [-inf,6.0) | -3 |
| [6.0,18.0) | -1 |
| [18.0,30.0) | 1 |
| [30.0,50.0) | 2 |
| [50.0,inf) | 6 |
| mths\_since\_recent\_inq | [-inf,3.0) | -8 |
| [3.0,5.0) | -3 |
| [5.0,7.0) | -1 |
| [7.0,inf) | 5 |
| mo\_sin\_old\_rev\_tl\_op | [-inf,60.0) | -15 |
| [60.0,80.0) | -10 |
| [80.0,140.0) | -4 |
| [140.0,170.0) | 1 |
| [170.0,inf) | 8 |
| income\_vs\_loan | [-inf,2.5) | -10 |
| [2.5,4.5) | -4 |
| [4.5,8.0) | 2 |
| [8.0,inf) | 8 |
| annual\_inc | [-inf,40000.0) | -7 |
| [40000.0,100000.0) | -1 |
| [100000.0,170000.0) | 6 |
| [170000.0,inf) | 8 |

**5.3. Hiệu quả mô hình**

|  |  |
| --- | --- |
| **GINI** | 36.72 |
| **KS** | 26.29 |