# Data exploration insights

### Overall

Enquiry Data happened before Account Data, and they can be merged by ‘customer\_no’. Each customer may have multiple records of enquiry and account. Timing difference between dates within this two files can be used to generate derived variables as same as numeric records like credit limit.

Each customer may have only one record in ‘raw\_data\_70\_new.csv’ as this file contains features of customer that may help predict his credit behavior. Some features like Date-of-birth and age represent the same characteristic, so either one is enough. Besides, features like tax number are not related with his credit behavior so may be dropped through IV or p-value during modeling phase.

### Account Data

There are two columns representing payment history, so while analyzing, we should combine them together while keep only the parts between quotation marks. As three character represent DPD in that very month, after combination, we also need to treat each three chars separately.

As some records with ‘opened\_dt’ lacking ‘last\_paymt\_dt’ but with ‘closed\_dt’, and ‘total\_diff\_lastpaymt\_opened\_dt’ should be greater for customer with worse credit record, when lacking ‘last\_paymt\_dt’, I set the ‘total\_diff\_lastpaymt\_opened\_dt’ to be the difference between ‘opened\_dt’ and ‘closed\_dt’, which also represents the extend of DPD.

### Enquiry Data

For Enquiry Data, only the relationship between ‘dt\_opened’ and ‘enquiry\_dt’ and the type of ‘enq\_purpose’ is utilized and the amount of enquiry is not. It might be reasoned from the fact that this feature represents the total amount in credit but acts less precise than quantitative value in Account Data file as it is the possible credit value instead of actual.

A mapping is needed to link code of enquiry purpose with secured / unsecured.

### Feature Data

Bad\_label being 0 or 1 represents the customer has good or bad credit record. Here we have 22892 lines of 0 and 1004 lines of 1.

Among all 79 features, 2 of them has only one level excluding missing value - feature\_5 is either 'Card Setup' or NA and feature\_6 is either 14 or NA; one of them is a duplicated of entry\_time which has already been considered in account data - feature\_2; redundant date-of-birth related features – feature\_21, feature\_63 and feature\_75; features with confused formats – feature\_70. All these abnormal features should be deleted before adding into the model.

Besides, as some features are read as character in R, we should reformat them into factor. NA’s are treating as level ‘level\_na’ before going further. As most of the factors are with more than five levels, we bin them first with ‘woeBinning’ package to reduce the number of variables.

As Xgboost can only deal with numeric variable, here we further transform the variable matrix into dummy variable matrix with binned levels.

# Feature matrix

The first 8 dummy variables selected with their gains are as below:

|  |  |  |
| --- | --- | --- |
|  | Feature | Gain |
| 1 | feature\_47.binned \_ xmisc. level pos. | 0.590317 |
| 2 | feature\_20.binned \_ xmisc. level pos. | 0.146516 |
| 3 | feature\_22.binned \_ xmisc. level pos. | 0.142252 |
| 4 | feature\_38.binned \_ xmisc. level pos. | 0.070428 |
| 5 | feature\_45.binned \_ xmisc. level pos. | 0.044747 |
| 6 | feature\_68.binned \_ x(1, Inf] | 0.001953 |
| 7 | feature\_77.binned \_ xmisc. level pos. | 0.001899 |
| 8 | feature\_16.binned \_ xmisc. level pos. + MA01 + AS17 | 0.001888 |

* *Gain: contribution of each feature to the model. For boosted tree model, each gain of each feature of each tree is taken into account, then average per feature to give a vision of the entire model. Highest percentage means important feature to predict the label used for the training (only available for tree models) – R*

# Model Evaluation

### Gini

Here I iterate for 50 times and have AUC of 0.989162, so

### Rank Ordering

|  |  |
| --- | --- |
| Decile Rank | Bad Credit Ratio |
| 10 | 0.909722222 |
| 8 | 0.67816092 |
| 6 | 0.510280374 |
| 5 | 0.147435897 |
| 4 | 0.172839506 |
| 3 | 0 |
| 2 | 0.058401639 |
| 1 | 0.000142952 |

* Note that here is no records available in ‘Decile Rank’ 9 and 7