# 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 features are read as character in R, we should decide whether to reformat them into numeric or factor. Features with less levels are transformed into factors and those quantitative value with large levels are transformed into numeric form.

# Feature matrix

### Selected by IV using ‘woeBinning’ Package

Information value for each variables are calculated and 50 variables are to be selected by setting ‘min.iv.total’ being the 50th greatest IV.

The first 10 variables selected with their IV are as below:

|  |  |
| --- | --- |
| Feature | IV |
| mean\_diff\_open\_enquiry\_dt | 12.73144 |
| ratio\_currbalance\_creditlimit | 11.36733 |
| feature\_20 | 10.49605 |
| mean\_diff\_lastpaymt\_opened\_dt | 10.38277 |
| total\_diff\_lastpaymt\_opened\_dt | 9.950712 |
| feature\_47 | 7.491298 |
| utilisation\_trend | 1.518612 |
| feature\_45 | 1.299038 |
| feature\_38 | 1.13217 |
| feature\_22 | 1.128002 |

Gains is not calculated as final model is not decided due to the failure of trying vif(), LASSO and step() and that the full model with all the IV-selected binned variables provides odds prediction being neither 0 or 1.

# Model Evaluation

### Gini

Sorts Bad\_Label according to their estimated possibility of falls into level 1 in descending order. We have 22892 good records and 1004 bad records, altogether 23896 in total.

Three lines are generated (X being the rank of sorted possibilities calculated by each model; Y being the percentage of actual bad record included in the records below (1 - (X / 23896)’th) quantile of possibility):

Optimal model which estimated ‘Bad\_label = 1’ with possibility 1 and ‘Bad\_label = 0’ with possibility 0 –

Generated Model:

Default Model:

Set the area between these three line and X-axis as 0pt, Gen and Def separately, we then have

### Rank Ordering

Sorts Bad\_Label according to their estimated ln(odds) of falls into level 1 in descending order.

Evenly separated the range of ln(odds) into 10 intervals and let the upper bound being x1, x2, … , x10.

The value represented as ’66 features (Ensemble)’ in the Test Questionnaire can be calculated by –

which represents the ratio of bad records against total records in each interval.