In a nutshell,

* HomeLoanDefaultRisk (ml\_train\_full): Base training data with all features generated
* HomeLoanDefaultRisk (ml\_test\_predictors): Base test data with all features generated
* HomeLoanDefaultRisk\_Sample (sample\_predictors01): Selected sample of ml\_train\_full (Target removed)
* HomeLoanDefaultRisk\_Sample (sample\_target): Selected sample of ml\_train\_full (Target only)
* HomeLoanDefaultRisk\_Sample (sample\_predictors01\_scaled\_ohe): on top of sample\_predictors01, missing values are filled with 0, numerical features are scaled, and categorical features are on-hot encoded. That is the final dataset I used to train.

Database 1 – Home Loan Default Risk

Table

* ml\_test\_predictors: Base test dataset after SQL Server feature engineering. It cannot be used directly in the model. It requires at least three additional steps (Usually done in R):
  + Feature selection (remove unused features such as SK\_ID\_CURR)
  + Fill missing values
  + One-hot encoding for categorical data
* ml\_train\_full: Base train dataset after SQL server feature engineering. It requires similar manipulation performed in the predictors dataset

Database 2 – HomeLoanDefaultRisk\_ES1

Table

* es1\_\*\*\*: dataset used for predicting missing external\_source01
* sd\*\*\*: source dataset directly from csv

Database 3 – HomeLoanDefaultRisk\_Sample

Table

* input\*\*\*: Input for R Machine Learning portal
* predictors01\_scaled\_ohe\_old: Previous version of training dataset (delegated)
* routput\*\*: R output data. Not useful
* sample\_base\_82k: Base sample data for training
* sample\_predictors01: Base training predictors dataset (still has missing value, not scaled, not one-hot encoded)
* sample\_predictors01\_scaled\_ohe: Final training data
* sample\_target: Final training target
* sample\_rfeed\*\*\*: delegated