# Input Space Experimentation for Logit Models Using Retail Mortgage Data

**Experiments**

* Aggregation of [g0\_Delinq]
  + Thrust: Aggregating the delinquency level variable may provide an indication of the level of default contagion, which other macro variables may fail to do. It can thus provide a “safety” net where other variables will fail to indicate stress within the portfolio. We identified two methods in which this aggregation can be done: (1) the proportion of accounts with any delinquency at each month (g0\_Delinq>0) (known as the “any” method), and (2) the proportion of accounts with zero delinquency (g0\_Delinq==0) and the proportion of accounts with one month’s delinquency (g0\_Delinq==1) (known as the “factorised” method); the second method uses two variables to capture the three possible levels of delinquency in performing accounts (g0\_Delinq==2 is implicit). The two variables are then lagged for 1-,2-,3-,4-,5-,6-,9-, and 12 months. The two versions of the variable along with their associated lags are tested separately by fitting logit models, running a best subset selection, and comparing the AUC values of the final model.
  + Setup: The two methods, as explained above, are used in aggregating the delinquency information across the portfolio. The three variables are then lagged for 1-,2-,3-,4-,5-,6-,9-, and 12 months. The two methods along with their associated lags are tested separately by fitting logit models, running a best subset selection, and comparing the AUC values of the final model.
  + Results: The “any” method results in a model with three variables and an AUC of 56.25%. The “factorised” method results in a model with 16 variables and an AUC of 56.52%. Although the “factorised” version results in a slightly better performance, it has more than 5 times the number of variables as the “any” method. It is thus suggested to use the “any” method with its associated lags; i.e., the current value and the 5-, and 12 months lags.
    - The 5- and 12 month lagged variables result in 6.5% and 2.6% of observations being omitted. This in itself is not too high, however the periods being omitted is in the start of an economic downturn (2007). It is thus suggested that only the 5 month lagged version be used, which along with the unlagged variable results in a model with an AUC of 55.8%.
* Aggregation of [InterestRate\_Margin]
  + Thrust: Aggregating this variable over each month in the sample window may give insight to the competitive pressures on the portfolio within the broader retail loans market. Since this variable has missing values, mean- or median imputation is needed, which necessitates a comparison between the two techniques for this variable. The question is then whether the aggregation should then be mean or median imputation.
  + Setup: Mean- and median imputation is first applied to [InterestRate\_Margin]. The two versions of the variable are then compared by fitting two logit models and comparing the AUC values. According to this test, the “best” imputation technique is chosen and aggregation is applied by taking the mean and median over each month in the sampling window. The two variables are then lagged for 1-,2-,3-,4-,5-,6-,9-, and 12 months. Finally, the two version of the variable along with their associated lags are tested separately by fitting logit models, running a best subset selection, and comparing the AUC values of the final model.
  + Results: Mean and median imputation to the interest rate margin variable result in logit models that have the exact same AUCs of 55.77%, there is thus no preference; mean imputation is chosen for the rest of the experiment. Mean aggregation to the interest rate margin model results in a best subset model with 3 variables and an AUC of 55.63%. Median aggregation to the interest rate margin model results in a best subset model with 6 variables and an AUC of 56.33%. It is suggested that median aggregation is used, with the unlagged variable and lags of 1-,2-,3-,9-, and 12 months.
    - The 9- and 12 month lagged variables result in 4.8% and 6.5% of observations being omitted. This in itself is excessively high, however the periods being omitted is in the start of an economic downturn (2007). It is thus suggested that only the 1-,2-, and 3 month lagged versions be used (these are the top three most important variables from the best subset model as identified by the custom function *varImport\_logit()*), which results in a model with an AUC of 56%.
* Aggregation of the number of new loans in the portfolio.
  + Thrust: The proportion of new loans being added to the portfolio each month may give an indication of the credit current appetite. This proportion can be computed using the number of new loans each month divided by the total number of active loans in that month, i.e., number weighting. Alternatively, it can be computed by dividing the total balance of the new loans in that month by the total balance of all active loans in that month, i.e., balance weighting. By number weighting we are only interested in the number of new cohorts entering the portfolio, whilst by balance weighting, we are interested in the total Rand exposure being introduced into the portfolio.
  + Setup: Two aggregation variables are first created which capture the proportion of new loans within each month of the sampling period, i.e., a number- and balance weighted variable. The two variables are then lagged for 1-,2-,3-,4-,5-,6-,9-, and 12 months. The two version of the variable along with their associated lags are tested separately by fitting logit models, running a best subset selection, and comparing the AUC values of the final model.
  + Results: Both versions of the aggregated variable results in a final model which includes the current aggregated value, and a 1-,3-,4-,5-, and 12 month lagged values. The three most important variables in descending order are the current value, the 12-month lagged value, and the 1 month lagged value. The AUC value of the number weighted model is slightly higher at 53.53%, whilst the weighted balance model has an AUC of 52.38%. Either version of the aggregated variable is suitable for use, but the number weighted model will has a slightly better performance.
    - The 12-,5-,4-,3-, and 1 month lagged variables result in 6.5%, 4.6%, 2.1%, 1.5%, and 0.5% of observations being omitted, respectively. This in itself is not too high, however the periods being omitted is in the start of an economic downturn (2007). It is thus suggested that the 1-,3-,4-, and 5 month lagged versions be used, which along with the unlagged variable results in a model with an AUC of 52.59%.
* Transformations
  + Thrust: Assessing the effect of transforming the input data used to train a logit model.
  + Setup: [Balance] and [M\_Repo\_Rate] are chosen as variables with which to apply transformations on and explore the effect of the transformations on the logit models. [Balance] is log-transformed and an optimal Yeo-Johnson transformation is applied. [M\_Repo\_Rate] is multiplied with 100 and squared and an optimal Yeo-Johnson transformation is applied. Three models, each containing the single raw- or transformed variable, are fitted and compared in terms of their AUC values for both [Balance] and [M\_Repo\_Rate]. The input space is then increased by adding the strong predictor [g0\_Delinq] and the three models are fitted and compared again.
  + Results: The AUCs are unaffected when transformations are applied and/or the input space is increased for either [Balance] and [M\_Repo\_Rate]. The residual deviance, AIC, and standard errors may however change when transformations are used. Further investigation may be necessitated.
* Loan-level delinquency volatilities/standard deviations
  + Thrust: The number of times which an account changes in delinquency may be an indication of the cohort’s financial stability. This stability may be captured by computing the standard deviation of the delinquency over various “windows”.
    - Too short windows, such as 2- or 3 months, may not be long enough to capture adequate information whilst too long windows, such as 9- or 12 months, may be too long as a lot of observations will need to be deleted due to the first n-1 periods not having any standard deviation values.
    - Note that a variable was created
  + Setup: SD variables are first created in the first data fusion script (script 2f.Data\_Fusion1) over various time windows, i.e., 4-,5-,6-,9-, and 12 months. Variables are tested by fitting logit models, running a best subset selection, and assessing the AUC value of the final model.
  + Results: The significant variables, ranked from most to least important, are the ones that compute the SD over 12-,6-,9-, and 4- months. The proportion of observations that are omitted when fitting the model is 13%, 10%, 6%, and 4% for the 4-,6-,9-, and 12 month SD variables, respectively. The 9- and 12 month variables were removed and the model refitted where the AUC dropped from 79.66% (for the best subset model) to 73.03% (for the reduced model). The trade-off between data omissions and the drop in AUC seems justifiable, it is suggested to use the loan-level delinquency SD variables over 4- and 6 month windows.

**Basic Variables**

* Application variables:
  + [Age\_Adj] vs [TimeInPerfSpell]
    - Thrust: Both variables are highly correlated from which a selection needs to be made.
    - Setup: Two models are fitted, i.e., one with the [Age\_Adj] variable and the second with the [TimeInPerfSpell] variable. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [Age\_Adj] has an AIC of 274583 and an AUC of 51.12%. The model with [TimeInPerfSpell] has an AIC of 274005 and an AUC of 60.11%. It is thus suggested to use [TimeInPerfSpell] over [Age\_Adj].
  + [Balance] vs [Instalment] vs [Principal]
    - Thrust: All three variables are highly correlated from which a selection needs to be made.
    - Setup: Three models are fitted, i.e., the first with [Balance], the second with [Instalment], and the third with [Principal]. The models are then assessed based on their AIC and AUC values.
      * In this special case, the correlation between them is also use in the final selection; see results below.
    - Results: The model with [Balance] has an AIC of 275014 and an AUC of 52.19%. The model with [Instalment] has an AIC of 274912 and an AUC of 53.47%. The model with [Principal] has an AIC of 273603 and an AUC of 57.5%. [Principal] is an account level variable (invariant over the history of an account), where as [Balance] and [Instalment] are period level variables (vary over the history of an account). Logically [Principal] would be kept and the choice is then between the two period level variables. [Balance] is kept since it has a lower correlation with [Principal] (80.45%) vs the correlation between [Instalment] and [Principal] (89.71%).
  + Testing two derivations of [InterestRate\_Margin].
    - Thrust: [InterestRate\_Margin] can be discritzed (using binning) which may be able to better discriminate between the risk profiles of accounts.
    - Setup: Testing is done post best subset selection of the application theme, i.e., the selection of variables from the best subset procedure is used where the [InterestRate\_Margin] variable is swapped between the two derivatives called [InterestRate\_Margin\_imputed\_mean] (the raw variable for whose missing values have been imputed using mean imputation) and [InterestMargin\_bin] (the binned version of [InterestRate\_Margin\_imputed\_mean]. The models are then compared based on their AIC and AUC values.
    - Results: The model with [InterestRate\_Margin\_imputed\_mean] has an AIC of 268583 and an AUC of 63.64%. The model with [InterestRate\_Margin\_bin] has an AIC of 268442 and an AUC of 63.39%. [InterestRate\_Margin\_imputed\_mean] is preferred since it has a slightly higher AIC value and results in a slightly more parsimonious model (compared to the binned version which needs two dummy variables); the AIC is slightly higher though (but not excessively so).
* Behavioural variables:
  + [slc\_acct\_prepaid\_perc\_dir\_12\_imputed] vs [slc\_acct\_pre\_lim\_perc\_imputed\_med]
    - Thrust: Both variables are highly correlated form which a selection needs to be made.
    - Setup: Two models are fitted, i.e., one with the [slc\_acct\_prepaid\_perc\_dir\_12\_imputed] variable and the second with the [slc\_acct\_pre\_lim\_perc\_imputed\_med] variable. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [slc\_acct\_prepaid\_perc\_dir\_12\_imputed] has an AIC of 275147 and an AUC of 61.37%. The model with [slc\_acct\_pre\_lim\_perc\_imputed\_med] has an AIC of 266557 and an AUC of 64.46%. It is thus suggested to use [slc\_acct\_pre\_lim\_perc\_imputed\_med] over [slc\_acct\_prepaid\_perc\_dir\_12\_imputed].
  + Derivation of payment method variable
    - Thrust: It is postulated that the “raw” payment variable [slc\_pmnt\_method] is too granular and that the number of levels of the variable introduce unnecessary complication. The variable is subsequently collapsed into [pmnt\_method\_grp].
    - Setup: Two models are fitted, i.e., one with the [slc\_pmnt\_method] variable and the second with the [pmnt\_method\_grp] variable. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [slc\_pmnt\_method] has an AIC of 252159 and an AUC of 73.06%. The model with [pmnt\_method\_grp] has an AIC of 254005 and an AUC of 72.73%. It is thus suggested to use [slc\_pmnt\_method] over [pmnt\_method\_grp].
  + Missing value indicators
    - Thrust: Some missing values in certain behavioural variables may be an indication of an account’s propensity to default.
    - Testing is done post best subset selection of the behavioural theme, i.e., the selection of variables from the best subset procedure is used. [slc\_acct\_pre\_lim\_perc\_imputed\_med] is allowed to interact with its missing value indicator variable [value\_ind\_slc\_acct\_pre\_lim\_perc]. Note that the “raw” variable is also kept in the model.
    - Results: The model with the missing value indicators produced coefficient estimates equal to “NA” for the interactions with the missing values. It is thus suggested that no missing value indicators are used in the behavioural variables.

**Intermediate Variables**

* Delinquency variables:
  + [PrevDefaults] vs [PerfSpell\_Num]
    - Thrust: Both variables are highly correlated, as expected since they give similar information (the one is just more granular than the other). A selection needs to subsequently be made.
    - Setup: Two models are fitted, i.e., one with the [PrevDefaults] variable and the second with the [PerfSpell\_Num] variable. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [PrevDefaults] has an AIC of 226134 and an AUC of 75.45%. The model with [PerfSpell\_Num] has an AIC of 263643 and an AUC of 60.04%. It is thus suggested to use [PrevDefaults] over [PerfSpell\_Num].
  + [g0\_Delinq\_Num] vs [PerfSpell\_g0\_Delinq\_Num]
    - Thrust: Both variables are highly correlated, as expected since they give similar information (the one is just more granular than the other). A selection needs to subsequently be made.
    - Setup: Two models are fitted, i.e., one with the [g0\_Delinq\_Num] variable and the second with the [PerfSpell\_g0\_Delinq\_Num] variable. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [g0\_Delinq\_Num] has an AIC of 25867 and an AUC of 77.24%. The model with [PerfSpell\_g0\_Delinq\_Num] has an AIC of 266003 and an AUC of 73.89%. It is thus suggested to use [g0\_Delinq\_Num] over [PerfSpell\_g0\_Delinq\_Num].
  + Missing value indicator for [slc\_acct\_roll\_ever\_24\_imputed\_mean]
    - Thrust: It can be suggested that missingness within this variable be an indication of some underlying mechanism that contributes to an account defaulting.
    - Setup: Two models are fitted, i.e., one with the [slc\_acct\_roll\_ever\_24\_imputed\_mean] variable and the second with the [slc\_acct\_roll\_ever\_24\_imputed\_mean] that is interacted with the associated missing value variable indicator [value\_ind\_slc\_acct\_roll\_ever\_24]; this latter model thus has two covariates, the raw variable and an interaction of the raw variable with the missing value indicator. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [slc\_acct\_roll\_ever\_24\_imputed\_mean] has an AIC of 234509 and an AUC of 78.61%. The model with [slc\_acct\_roll\_ever\_24\_imputed\_mean] and [value\_ind\_slc\_acct\_roll\_ever\_24] has an AIC of 233243 and an AUC of 78.14%. It is thus suggested to use [slc\_acct\_roll\_ever\_24\_imputed\_mean] only.
  + Missing value indicator for [slc\_past\_due\_amt\_imputed\_med]
    - Thrust: It can be suggested that missingness within this variable be an indication of some underlying mechanism that contributes to an account defaulting.
    - Setup: Two models are fitted, i.e., one with the [slc\_past\_due\_amt\_imputed\_med] variable and the second with the [slc\_past\_due\_amt\_imputed\_med] that is interacted with the associated missing value variable indicator [value\_ind\_ slc\_past\_due\_amt]; this latter model thus has two covariates, the raw variable and an interaction of the raw variable with the missing value indicator. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [slc\_past\_due\_amt\_imputed\_med] and [value\_ind\_ slc\_past\_due\_amt\_imputed] has an estimated coefficient of “NA”. Therefore, use only the raw variable [slc\_past\_due\_amt\_imputed\_med].
  + Non-linear relationship between delinquency and default.
    - Thrust: Anecdotal evidence from other projects suggest that there exists a non-linear relationship between delinquency ([g0\_Delinq]) and default.
    - Setup: Two models are fitted, i.e., one with the [g0\_Delinq] (numeric) variable and the second with the [g0\_Delinq\_fac] (factor) variable. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [g0\_Delinq] has an AIC of 215934 and an AUC of 74.03 %. The model with [g0\_Delinq\_fac] has an AIC of 215115 and an AUC of 74.03%. It is thus suggested to use [g0\_Delinq] over [g0\_Delinq\_bin]; there is no strong evidence to suggest that the factorised variable produces a superior model (it is also less parsimonious since two dummy variables are required).

**Advanced Variables**

* Portfolio level information (aggregation of variables)
  + [AgeToTerm\_Aggr\_Mean] vs [PerfSpell\_Maturity\_Aggr\_Mean]
    - Thrust: Both variables are highly correlated form which a selection needs to be made.
    - Setup: Two models are fitted, i.e., one with the [AgeToTerm\_Aggr\_Mean] variable and the second with the [PerfSpell\_Maturity\_Aggr\_Mean] variable. The models are then assessed based on their AIC and AUC values.
    - Results: The model with [AgeToTerm\_Aggr\_Mean] has an AIC of 272048 and an AUC of 58.10%. The model with [PerfSpell\_Maturity\_Aggr\_Mean] has an AIC of 274823 and an AUC of 51.95 %. It is thus suggested to use [AgeToTerm\_Aggr\_Mean] over [PerfSpell\_Maturity\_Aggr\_Mean].