|  |  |
| --- | --- |
| Modeling Team  Risk Controlling Department  Risk Management Division |  |

**COLLECTION MODEL FOR B03 RETAIL UNSECURED AND CREDIT CARD**

**Model development report**

Modelling Team

October, 2023

|  |  |
| --- | --- |
| Author: | Tran Quoc Thai |
| Contact: | 0342540765 |
| Creation Date: | 03/08/2023 |
| Last Modification Date: | 16/10/2023 |
| Version: | 1 |
| Status: | Final Version |

**Document Control**

**Change Record**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Author** | **Version** | **Change Reference** |
| 16/10/2023 | Tran Quoc Thai | 1.0 | Final version |
|  |  |  |  |

Table of content

[1. Executive summary 6](#_Toc138320666)

[2. Model design & Model scope 7](#_Toc138320667)

[2.1. Model design 7](#_Toc138320668)

[2.2. Model scope 7](#_Toc138320669)

[3. Model development methodology 8](#_Toc138320670)

[3.1. Single factor analysis 8](#_Toc138320671)

[3.1.1. WOE transformation 8](#_Toc138320672)

[3.1.2. Information value 10](#_Toc138320673)

[3.1.3. Binning method 10](#_Toc138320674)

[3.2. Correlation analysis 11](#_Toc138320675)

[3.2.1. Clustering analysis 12](#_Toc138320676)

[3.3. Logistic regression 13](#_Toc138320677)

[3.3.1. Effects-Selection Method 14](#_Toc138320678)

[3.3.2. Marginal IV. 14](#_Toc138320679)

[3.4. Performance measure 15](#_Toc138320680)

[3.4.1. Discrimination 15](#_Toc138320681)

[3.4.2. Accuracy 17](#_Toc138320682)

[3.4.3. Stability 18](#_Toc138320683)

[3.5. Model calibration 19](#_Toc138320684)

[3.5.1. Offset calibration 19](#_Toc138320685)

[3.5.2. Target calibration 20](#_Toc138320686)

[3.5.3. Score band calibration 20](#_Toc138320687)

[3.6. Scaling calculation 21](#_Toc138320688)

[4. Development process 21](#_Toc138320689)

[4.1. Target definition 22](#_Toc138320690)

[4.1.1. Detail definition of target 22](#_Toc138320691)

[4.1.2. Confirm the target definition 22](#_Toc138320692)

[4.2. Data preparation 24](#_Toc138320693)

[4.2.1. Data source 24](#_Toc138320694)

[4.2.2. Sample selection 25](#_Toc138320695)

[4.2.3. Data exclusion 26](#_Toc138320696)

[4.2.4. Explanatory variable 27](#_Toc138320697)

[4.2.5. Data quality 27](#_Toc138320698)

[4.3. Segmentation 28](#_Toc138320699)

[4.4. Single factor analysis 28](#_Toc138320700)

[4.5. Multi factor analysis 29](#_Toc138320701)

[4.6. Logistic regression 29](#_Toc138320702)

[4.6.1. Result of logistic regression in SAS 29](#_Toc138320703)

[4.6.2. List of variables in model 30](#_Toc138320704)

[4.6.3. Explain variables in the model 30](#_Toc138320705)

[4.7. Calibration/transformation 36](#_Toc138320706)

[4.8. Model statistic assessment 39](#_Toc138320707)

[4.8.1. Model discrimination 39](#_Toc138320708)

[4.8.2. Model accuracy 40](#_Toc138320709)

[5. Pre-validation 40](#_Toc138320710)

[5.1. Model discrimination 40](#_Toc138320711)

[5.2. Model accuracy 40](#_Toc138320712)

[5.3. Model stability 40](#_Toc138320713)

[5.4. Model weakness/limitation/assumption 41](#_Toc138320714)

[5.5. Plan of improvement 41](#_Toc138320715)

[6. Model usage 41](#_Toc138320716)

[6.1. Model implementation 41](#_Toc138320717)

[6.2. Responsibility of related party 41](#_Toc138320718)

[6.3. Related policy 41](#_Toc138320719)

[6.4. Monitoring report 41](#_Toc138320720)

[7. Model makers 42](#_Toc138320721)

[8. Appendix 42](#_Toc138320722)

[9. Bibliography 42](#_Toc138320723)

**List of tables**

[Table 1: Gini and PSI of model 6](#_Toc28162550)

[Table 2: Sample overview 7](#_Toc28162551)

[Table 3: Data filter 8](#_Toc28162552)

[Table 4: List of participants 37](#_Toc28162553)

**List of figures**

[Figure 1: Sampling methods 11](file:///G:\01.MODEL\II.%20MODEL_2019\01.20190125_UPL_V5\09.%20FINAL\DOCUMENTATION\Model%20development%20documentation.docx#_Toc28162296)

[Figure 2: Stratified Sampling 11](file:///G:\01.MODEL\II.%20MODEL_2019\01.20190125_UPL_V5\09.%20FINAL\DOCUMENTATION\Model%20development%20documentation.docx#_Toc28162297)

[Table 3: Hierarchical clusters tree 17](#_Toc28162298)

**Abbreviations:**

Finnone: Loan original system, which was used in VPBank from 2013 to 2017.

RLOS: Retail loan original system, which replace Finnone from 2016.

DWH: Data warehouse.

PCB: Vietnam credit information joint stock Company.

CSR: Credit support representative.

VPBank: Vietnam Prosperity Joint-Stock Commercial Bank.

# Executive summary

|  |  |  |  |
| --- | --- | --- | --- |
| **General Information** | | | |
| Date of model development | The model was developed in the period from September 2023 to October 2023 | | |
| Title | Collection model for B03 retail unsecured and credit card | | |
| Version | 1.0 (First generation) | | |
| Portfolio for model | Retail | | |
| List of systems on which the model is implemented | The model will be implemented in the RLOS systems.  RLOS systems:   * keeps information on application form and other information of borrowers; * determines borrower's score; * …   The system is managed by IT department. | | |
| Date of implementation |  | | |
| Brief description of the model | Logit regression method was used in the model to predict probability of customers will be over due 90+ in next 2 months for customer | | |
| Rating scale | <Mô tả rating scale dùng cho mô hình. Đối với các mô hình chưa có rating scale mà chỉ dừng lại ở scoreband, chia màu (eg. Mô hình collection) thì mô tả scoreband và cách chia màu>  Master scale is created separately for the Bank’s needs, and currently applied to the model – consists of 10 non-default grades (where each grade has 2 notches), and 1 default grade. | | |
| Departments which employ the rating model | Retail Risk | | |
| Approved by | Scoring committee (CRO, CEO) | | |
| Date of approval |  | | |
| Head of department responsible for the model | Minh Chu Hong (RMD - CrMD) <minhch@vpbank.com.vn> | | |
| Employee responsible for the model | Tran Quoc Thai (RMD - RMOD) <thaitq15@vpbank.com.vn> | | |
| Portfolio statistics **as of 04.08.2023 <thời điểm xây dựng mô hình>** | |  | |
| Number of clients | 12778 | | |
| Loan portfolio amount |  | | |
| Development data statistics (mẫu xây dựng mô hình) | |  | |
| Development sample | 02.2023 – 04.2023 | | |
| Good / bad / total observations | 1 344/11 434/12 778 | | |
| Gini on development sample | 62.54% | | |
| In time validation data statistics (mẫu 30% - nếu có) | | | |
| In time validation sample | 02.2023 – 04.2023 | | |
| Good / bad / total observations | 575 / 4899 / 5474 | | |  |
| Gini on development sample | 61.19% | | |  |
| Initial validation data statistics (mẫu OOT) | | | |
| Validation sample | 05.2023 – 06.2023 | | |
| Good / bad / total observations | 1 212 / 12 368 / 13 580 | |  |
| Gini on validation sample | 60.13% | |  |
| Stability analysis | | | |
| Recent sample | 04.2023 – 07.2023 | |  |
| PSI on recent sample | 6.37% | |  |

Table : Executive summary

# Model design & Model scope

Model was built using logistic regression in which variables were analyzed and assessed through WOE and IV. Benefits of this method:

* Reduce impact of outlier and missing observation;
* Variables chosen into model have predictive power as well as concurrent with reality;
* Easy to implement in scoring system;
* Easy to control stability and predictive power after development.

Model development steps are presented in following sections.

## Model design

Theo model is built for Customers use Unsecured/OD/Credit card, belong to bucket B03 (DPD 60-90), not exist in Badbank and are not partner (Deposit home loans, Bond loans, Deaura, Kiman products) customers pay 100% when overdue.

## Model scope

<mô tả portfolio mà mô hình đang được thiết kế để sử dụng, bao gồm ít nhất các thông tin sau:

Đối với application model/rating

* Danh mục tại thời điểm xây dựng mô hình: volume, tốc độ phát triển, NPL
* Đặc trưng của danh mục/KH
* Nhóm khác hàng mục tiêu, target của năm tới.
* Cách thức xác định danh mục (dựa theo sizing của KH hay theo sản phẩm, tài sản, …), độ ổn định về cách xác định danh mục.
* Trong trường hợp KH đổi khỏi danh mục đó thì rating được thực hiện ntn

Đối với các mô hình khác:

* Danh mục tại thời điểm xây dựng mô hình: volume, tốc độ phát triển, NPL (chỉ số NPL có thể đổi thành các chỉ số khác phù hợp với việc sử dụng mô hình)
* Đặc trưng của danh mục/KH
* Cách thức xác định danh mục (dựa theo sizing của KH hay theo sản phẩm, tài sản, …), độ ổn định về cách xác định danh mục.
* Trong trường hợp KH đổi khỏi danh mục đó thì việc scoring được thực hiện ntn (không cần scoring hay thuộc phạm vi của mô hình khác)>

# Model development methodology

Model was built using logistic regression in which variables were analyzed and assessed through WOE and IV. Benefits of this method are showed below

* Reduce impact of outlier and missing observation;
* Variables chosen into model have predictive power as well as consistent with business sense;
* Easy to implement in scoring system;
* Easy to control stability and predictive power after development.

Model development steps are presented in following sections.

## Single factor analysis

The single factor analysis is done using WOE transformation and IV assessment.

### WOE transformation

The general variable analysis method composes of:

* Dividing variable into different small groups (e.g., 20 groups), calculating WOE for each group.
* Appropriately combining the small groups with similar properties/ WOE values, edit the cut points (if needed) to fit the reality as well as logical trend.
* A general “minimum 5 percent in each bucket” rule has been applied to enable meaningful analysis. There are also a sufficiently high number of good and bad cases in each bucket. At minimum, industry practitioners look for a minimum of 80 to 100 cases in each bin, but this number may be higher when dealing with larger data sets. However, the extreme groups (the worst or the best) can be accepted less than 5% as long as the number of observations is large and there is a reasonable explanation.
* There are no groups with 0 counts for good or bad. When using auto binning algorithms on low default portfolios, if a bin is formed with 0 goods or bads, analysts normally assume a small number (1 or 0.5) of goods or bads in order to calculate the WOE.
* The bad rate and WOE are sufficiently different from one group to the next (i.e., the grouping has been done in a way to maximize differentiation between goods and bads, and from one group to the next). This is one of the objectives of this exercise to identify and separate attributes that differentiate well. While the absolute value of the WOE is important, the difference in WOE between the groups is key to establishing differentiation. The larger the difference between subsequent groups, the higher the rank ordering ability of this characteristic.
* The WOE for non-missing values also follows a logical distribution, going from negative to positive without any reversals. This confirms business logic

After completing the grouping process, the WOE values are calculated for each group and variables are transformed in WOE form before estimate coefficients. Formula of WOE is calculated as follow

In which:

* is percentage of Good observation in each group divided by total Good obseravtion in sample.
* is percentage of Bad observation in each group divided by total Bad observation in sample.

Models can also be created using continuous (ungrouped) characteristics however, the grouping process offers some advantages:

* It offers an easier way to deal with outliers with interval variables, and rare classes.
* The grouping process makes it easy to understand the relationship between predictor and dependent variable and therefore gain more knowledge of the portfolio.
* Nonlinear dependencies can be modeled with linear models.
* It helps to reduce the degree of freedom of the variable (when compare to dummy transformation).
* The variables got from WOE transformation have the similar range.
* The relationship between predictors and target are easy to present to business.

### Information value

Information Value (IV) indicator is used to assess variable’s ability to classify Good/Bad in dependent variable. It is calculated as follow:

Variables with low IV value have weak classification capability. The variables will be excluded after analysis. The benchmarks for IV are as follow:

* : no classification capability.
* : average classification capability.
* : good classification capability.
* : excellent classification capability.
* : tremendous classification power, can be considered to make policies (in application models).

### Binning method

Combination (binning) is the process of moving from coarse binning ( groups) to fine binning ( groups, *constraint*). The combination step shall be done in two ways:

#### Manual binning

Manual binning can be done as follows:

* Merge WOE groups,
* Merge based on expected trend of variables.
* Meet the standards in the WOE transformation section.

#### Auto-binning

At each number, the algorithm will find all possible ways to divide groups into sub-groups and calculate IV for each way. The number of ways is:

Next, the algorithm will find a way with highest IV and satisfy constrain. The example result was shown as below:

The number is chosen by the following ways:

* Observe the IV by the number of group line to balance between forecasting and complexity.
* Satisfying constraint: (e.g. monotonic ...)

## Correlation analysis

#### Pearson correlation

After single factor analysis, Pearson correlation (formula shown below) is calculated to choose variables for running logistic regression.

Chosen variables must have correlation with no more than 0.5 (with application model) and 0.7 (with behavior model).

#### VIF Test

In statistics, the variance inflation factor (VIF) is the ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone. It quantifies the severity of multi-collinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

Modelling team can calculate different VIFs (one for each ) in three steps:

**Step 1.** Run an ordinary least square regression that has as a function of all the other explanatory variables in the first equation. If , for example, equation would be

where is a constant and is the error term.

**Step 2.** Calculate the VIF factor for with the following formula:

where is the coefficient of determination of the regression equation in step one, with on the left hand side, and all other predictor variables (all the others) on the right hand side.

**Step 3.** Analyze the magnitude of multicollinearity by considering the size of the . A rule of thumb is that if then multicollinearity is high (a cutoff of 5 is also commonly used).

### Clustering analysis

The clustering analysis divides a set of numeric variables into disjoint or hierarchical clusters. These analyses use a type of principal components analysis to identify groups of characteristics that are correlated. One can then select one or more characteristics from each group, and theoretically, represent all the information contained in the other characteristics in each of the groups. In addition, business considerations should also be used in selecting variables from this exercise, so that the final variables chosen are consistent with business preference

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables 1 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 2 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 3 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 4 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 5 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 6 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 7 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 8 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 9 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 10 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 11 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 12 |  |  |  |  |  |
|  |  |  |  |  |
| Variables 13 |  |  |  |  |  |
|  |  |  |  |  |

Table 7: Hierarchical clusters tree

## Logistic regression

Logistic regression analysis is often used to investigate the relationship between these discrete responses and a set of explanatory variables. Texts that discuss logistic regression include Agresti (2013); Allison (2012); Collett (2003); Cox and Snell (1989); Hosmer and Lemeshow (2013); Stokes, Davis, and Koch (2012).

For binary response models, the response, , of an individual or an experimental unit can take on one of two possible values, denoted for convenience by 0 and 1 (for example, if a customer is bad, otherwise ). Suppose is a vector of explanatory variables and is the response probability to be modeled. The linear logistic model has the form

where is the intercept parameter and is the vector of s slope parameters.

The LOGISTIC procedure fits linear logistic regression models for discrete response data by the method of maximum likelihood.

The maximum likelihood estimation is carried out with either the Fisher scoring algorithm or the Newton-Raphson algorithm

### Effects-Selection Method

Logistic regression is used with three types of technique to estimate which variables will enter the model:

#### *Forward selection*

First selects the best one characteristic model based on the individual predictive power of each characteristic, then adds further characteristics to this model to create the best two, three, four, and so on characteristic models incrementally, until no remaining characteristics have p-values of less than some significant level (e.g., 0.5), or univariate chi-square/minimum discrimination information statistic above a determined level. This method is efficient, but can be weak if there are too many characteristics or high correlation. This method can however be modified for business usage, as we will do in the next section.

#### Backward elimination

The opposite of forward selection, this method starts with all the characteristics in the model and sequentially eliminates characteristics that are considered the least significant, given the other characteristics in the model, until all the remaining characteristics have a p-value below a significant level (e.g., 0.1) or based on some other measure of multivariate significance. This method allows variables of lower significance a higher chance to enter the model, much more than forward or stepwise, whereby one or two powerful variables can dominate.

#### Stepwise

A combination of the preceding two techniques, this involves adding and removing characteristics dynamically from the scorecard in each step until the best combination is reached. A user can set minimum p-values (or chi-square) required to be added to the model or to be kept in the model.

### Marginal IV.

Select variables by Marginal IV by following steps:

1. At initial step, set .
2. Calculate Marginal IV for all variables according to the formula:

Where:

* is number of good/bad in attribute.
* is sum of predicted good/bad in attribute.

1. Select the variable with the highest Marginal IV into the model and recalculate the score.
2. Recalculate the marginal IV according to the score in step 3 (note that the variable that was included in the model has a marginal IV of 0).
3. Back to step 1

The current parameters are **1) stepwise method, 2) significant level for entry is 0.05, 3) significant level for stay is 0.05 and 4) maximum correlation of variables in the model**.

## Performance measure

### Discrimination

#### Gini Coefficient

When we plot cumulative score distribution from the goods against the cumulative score distribution from the bad cases, we get what is known as the Trade-off curve. For each score distribution, the plotted trade-off curve shows how many Bad customers are captured by the score at the low end of the distribution for a given number of good records. In general, the farther the score trade-off curve bows to the left and up, the better the scoring model. The diagonal 45⁰ line is the “line of random selection” and is the equivalent of a scoring system that does not work at all. The Gini Coefficient is defined as the ratio of area between the Trade-Off curve and the diagonal line to the area above the diagonal line.

The benchmark for Gini Coefficient:

|  |  |  |
| --- | --- | --- |
| **Behavior Model** | **Application Model** | **Benchmark** |
| Gini<50 | Gini<30 | Weak |
| 50<=Gini<60 | 30<=Gini<40 | Medium |
| 60<=Gini | 40<= Gini | Strong |

#### ROC Area (ROC)

Similar to calculation of the Gini Coefficient, the ROC is determined from the cumulative score distribution from the goods against the cumulative score distribution from the bad cases, known as the Trade-off curve. However, the ROC is the unscaled area under the Trade-Off Curve, instead of the value scaled by subtracting and dividing by 0.5

#### K-S Statistic

K-S is a non-parametric measure. It determines whether two underlying distributions differ by comparing their cumulative distributions. It is defined as the maximum vertical distance between the two cumulative distributions.

The formula for K-S calculation:

Where and are the cumulative score distribution functions of goods and bad cases respectively.

### Accuracy

Modelling team compared the predicted bad rates with the actual bad rates and tested the model performance in terms of accuracy using Mean absolute percentage error (MAPE).

MAPE is a measure of prediction accuracy in a trend estimation. The difference between actual and predicted is divided by the actual value. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. Multiplying by 100 makes it a percentage error.

This expresses accuracy as a percentage of the actual and hence can be used as to measure closeness of predicted to observe in any model. The formula for MAPE is:

Where

* is actual bad rate in group ,
* is predicted bad rate in score band ,
* is number of score bands

Indicative Tolerance

* signifies strong closeness of prediction to actual.
* signifies moderate closeness of prediction to actual.
* signifies poor closeness of prediction to actual.

### Stability

It is vital that the model developed is built on a population that has remained consistent over a period of time. Stability is measured by the Population Stability Index (PSI). It is assessed in two ways; stability at the characteristic level and stability of the entire population at totality level. The purpose of this validation is to assess the stability of the overall distribution by comparing the sample data in the development period to the most recent year of data available, which reflects the scoring population in the live environment.

PSI is computed using the formula shown below by aggregating the shift from all deciles of the population score distribution. Characteristic Stability Index (CSI) applies the same formula, by aggregating from all the bins of the characteristic.

Where

* : Ratio of observations in baseline sample for decile
* : Ratio of observations in validation sample

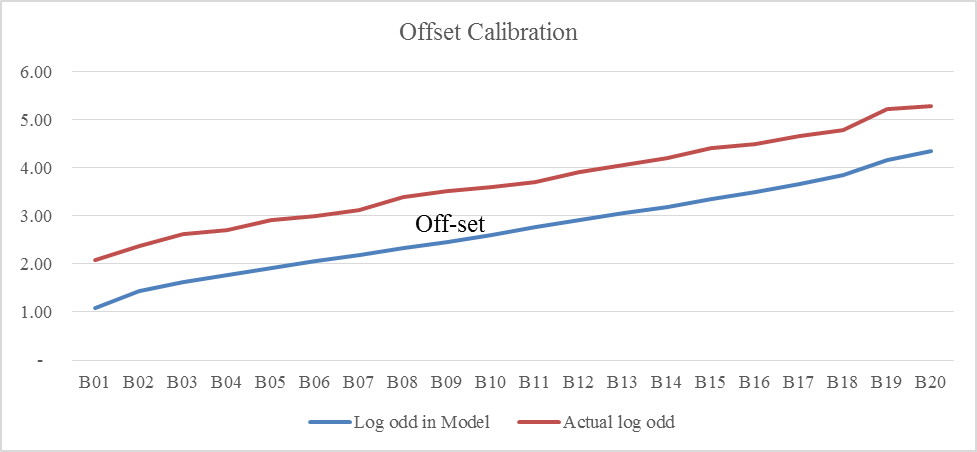
Below does Modelling Team recommend the benchmark of PSI:

* : Model is stable.
* : Model is moderately stable.
* : Model is unstable.

## Model calibration

### Offset calibration

Offset calibration method assumes the log odd of the actual sample is less/more than the log odd of the development sample a constant value. The method bases on Bayes’s theorem about the probability of an event, based on prior knowledge of conditions that might be related to the event.



The calibration process is described below:

* Build logit regression model:
* Calculate overall bad rate of development sample (p).
* Calculate overall bad rate of actual sample (P).
* Calculate off-set:
* Calculate calibrated probability:

### Target calibration

Target calibration calculates the calibrate coefficient (scalar) so that the average of the predicted values is equal to the bad rate of actual sample (P). The steps are as follows:

* Build logit regression model:
* Calculate overall bad rate of actual sample (P).
* Solve for such that:
* Calculate calibrated probability:

### Score band calibration

Offset calibration method assumes log odd of the actual sample has correlation linear with the log of the development sample. The method bases on the optimal calculation to log odd of development sample to fit with actual log odd.

The calibration process is described below:

* Score for actual sample.
* Divide the samples into groups (score bands) based on predicted log odd (usually 20 groups).
* Calculate average log odd of each group .
* Calculate actual log odd of each group:
* Run OLS regression:
* Calculate calibrated probability:

where A is slope and B is intercept from OLS model.

## Scaling calculation

In order to implement model in system, model need to be scale into score. Scaling refers to the range and format of scores in a scorecard and the rate of change in odds for increases in score. The choice of scaling, or its parameters, does not affect the predictive strength of the scorecard. The formula below shows the relationship between odds and scores:

Modelling team refer two references points, **where at a score of 500 the odds are 2.25 and at a score of 600, the odds should be 5**.

The formula below shows scores calculation for each attribute:

Where

* is score for attribute of variables
* is parameter for variable

The intercept is considered separately:

# Development process

## Target definition

### Detail definition of target

The Good/Bad/Indeterminate is defined as following:

* Good observation: keep and move to lower bucket (keep or Lower than B03) at the end of each month in outcome period
* Bad observations: Move to higher bucket (Greater than B03) at the end of each month in outcome period

According to the above definition, the number of Good and Bad observations in the sample are as follow.

|  |  |  |
| --- | --- | --- |
|  | **Frequency** | **Percent** |
| **Bad:** Move to higher bucket (Greater than B03) at the end of each month in outcome period | 11 434 | 89.5% |
| **Good**: keep and move to lower bucket (keep or Lower than B03) at the end of each month in outcome period | 1 344 | 10.5% |
| **TOTAL** | **12 778** | **100%** |

Table : Training sample overview

Performance window was chosen with 2 months from the disbursement date.

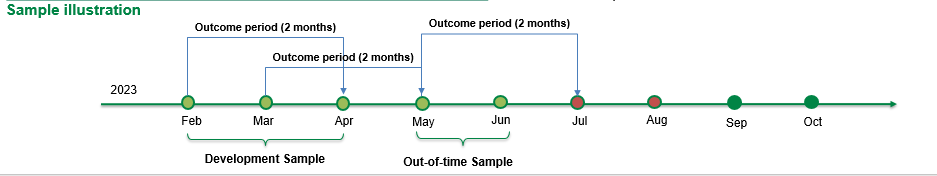


Figure : Observation window and performance window

### Confirm the target definition

## Data preparation

### Data source

In order to comprehensively assess the risk aspects of the customer, four sources of information were collected as follows:

* Application information: is information about the demographics of borrowers, number of rejected applications, approved applications, liabilities, time in relationship with VPBank, CIC information, collateral information... These information are in loan application form and other documents provided by customers. These information are inputted in RLOS system by CRS and consolidated in DWH by IT. Modelling team queries data directly from DWH.
* Loan information: is information about number of loans, type of loan, historical repayment, paid amount… These information are taken from the source tables on DWH. The Modelling team queries data directly from DWH.
* PCB Information: is information about credit histories of customers at other credit institutions (except VPBank), including information about loans, credit cards, debt groups, etc. This information is collected, calculated and returned by the PCB and IT to Modelling team at the time of model development.
* Card, deposit and current account:  Information about behavior of borrowers including information of deposits, number of current account transactions (in and out), amount of transactions… These data are taken from the source tables on DWH.

Details of data sources are shown in the table below:

|  |  |
| --- | --- |
| Group - variable | Source |
| Demographic variables | [10.36.31.16]. VPB\_WHR2.dbo.TBL\_RLOS\_COMM\_APP |
| [10.36.31.16]. VPB\_WHR2.dbo.TBL\_RLOS\_COMM\_APP\_FORM |
| [10.36.31.16].VPB\_WHR2.dbo.TBL\_RLOS\_APPLICANT\_RELATION\_INFO |
| [10.36.31.16].VPB\_WHR2.dbo.T24\_COLLATERAL\_ALL |
| Loan | [10.36.31.16].VPB\_WHR2.dbo.T24CRD |
| [10.36.31.16].VPB\_WHR2.dbo.T24CRD\_LD\_SCHEDULE\_NEW |
| PCB | [10.36.31.16].VPB\_WHR2.dbo.TBL\_RLOS\_APP\_PCB\_WIDEGT |
| Card, deposit and current account | [10.36.13.74].BICDATA.dbo.CARD\_ENDMONTH |
| [10.36.31.16].VPB\_WHR2.dbo.TBL\_W4\_CONTRACT |
| [10.36.31.16].VPB\_WHR2.dbo.FT\_HIST |
| [10.36.31.16].VPB\_WHR2.dbo.TT\_HIST |
| [10.36.13.74].BICDATA.dbo.LNTB\_MASTER\_ENDMONTH |
| Other | [10.36.31.16].VPB\_WHR2. dbo.VPB\_CUSTOMER |

Table : Data source

### Sample selection

The model includes 20282 B03 Unsecured/OD/Credit card from February 2023 to April 2023. After data filtering, the data in the model was composed of 18252 customers. Details of the filter process will be presented later.

|  |  |  |
| --- | --- | --- |
| **Data** | **Time frame** | **#Observations** |
| Development | 02.2023-04.2023 | 18252 |
| Out of time validation | 05.2023-06.2023 | 13580 |

Table 3: Sample overview

In order to determinate sample window, we made analysis of bad rate over than one year history data:

Figure : Bad rate analysis

As can be seen from above chart, bad rate has no significant change from February 2023. In addition, in order to have enough time to identify good / bad, the time of development sample needs to be back in the past at least 3 months from the time of building the model. Therefore, the period from February 2023 to April 2023 is proposal sample window.

### Data exclusion

Data filter is performed in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Sample** | **# Observations** | **%Fitter** |
| 1 | B03 unsecured, OD, credit card product customer | 20282 | 100% |
| 2 | Remove observations which belong to Bad bank list | 20136 | 99.3% |
| 3 | Remove Roll Back T-1 | 18252 | 89.9% |

Table 4: Data filter

Step 1: All loans disbursed from 02.2023 to 04.2023 B03 unsecured, OD, credit card product customer

Step 2: Remove observations which belong to Bad bank list

Step 3: Remove Roll Back T-1

### Variable description

|  |  |
| --- | --- |
| Group | Detail |
| Loan | * Historical delinquency * Loan information, variation in balance, etc. |
| Collateral & Deposit | * Number of collateral * Total deposit amount |
| Current account | * Number of spend count * Payment count (in and out) |
| Credit card | * Historical delinquency * Credit card transaction * Credit card transaction, utilization, etc. |
| Historical Application | * Historical application * Demographic information |
| Collection | * Number of action, % of action type * Response of customer, etc * Indicator from Retail Risk |
| Credit bureau | * PCB information * CIC information |
| Loan + Credit card | * Historical delinquency * Loan and credit card information, variation in balance, etc |
| Others | * Cross variables |

### Data quality

According to data source and filtering process described above, the data used to build the model has the following problems:

Demographic data:

* Data are lack of trust: demographic information is collected from customer’s application form and it is entered by CSR on the system. Some of the information does not have verification papers. For example, education, working time in the current company, time in current address. In addition, CRS sometimes doesn’t input correctly data into system, for example, some differences between data in RLOS and hard copy are found when modeling team checking outlier value of data.
* Data missing for not mandatory fields: some not-mandatory fields have high portion of missing value (> 50%), for example: working time in the previous company, number of employees…

Internal behavior data:

* Data missing: Data used to build models is in the New to Bank segment; therefore, most of them do not have credit relationship with VPBank. The missing percentage of these variables is approximately 90%. Thus, behavior variables have weak predictive power although they are the most reliable.
* Requirement for fast implementation in RLOS: in order to implement model in RLOS, behavior information need to be ready in RLOS’s datamart. Thus, in development sample, only available variables were considered to facilitate model implementation. This makes limit for number of behavior variables in development. However, it is not affected significant to model result because the fact mentioned above about data missing.

PCB data:

* Missing data: the percentage of applications that have information in the PCB data warehouse reaches over 40%. This hit rate is quite good, many PCB variables have ability to predict.
* Data are collected, calculated and extracted by PCB to VPBank so Modeling team can not verify accuracy of the data.

Detail % of missing for each variable is presented in the sheet “List variable” in file “Final\_upl\_pcb\_2018.xlsx” attached in appendix. Data workflow

## Single factor analysis

Before being analyzed, data is processed through the following steps:

* Handling missing: Almost missing in variables are meaningful => Can’t be removed
* Check the plausibility of the variables with time: 1 month, 3 months, 6 months, 9 months and 12 months.
* Evaluate the suitability and characteristic of each variables

The results of data analysis are shown in the table below:

|  |  |
| --- | --- |
| Criteria | No. varibales |
| All single variables | 849 |
| Number of variables with a small IV (IV<0.02) -> Remove | 690 |
| Number of variables trend to be illogical -> Remove | 270 |
| Number of variables missing > 98% or unmeaningful misssing -> Remove | 260 |
| Number of variables used in logistic regression | 249 |

## Multi factor analysis

List of variables used for logistic regression satisfying the following conditions:

* IV> = 0.02
* Trend of the variables is reasonable according to the business.
* The trend of variables doesn’t change in the out of time sample.

## Logistic regression

The criteria are given for selecting variables in the model:

* Remove variables with negative coefficients.
* Remove variable has a correlation variable higher than 0.7 or CSI high.

### Result of logistic regression in SAS

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Analysis of Maximum Likelihood Estimates** | | | | | | | | | |
| **Parameter** | | **DF** | | **Estimate** | | **Standard** | | **Wald** | **Pr > ChiSq** |
|  | |  | |  | | **Error** | | **Chi-Square** |  |
| Intercept | | 1 | | -2.1363 | | 0.0341 | | 3920.995 | <.0001 |
| WOE\_N\_DPD\_EOM\_CHANGE\_M3 | | 1 | | 0.2693 | | 0.0489 | | 30.3518 | <.0001 |
| WOE\_N\_LOWDPD\_ALL\_C12 | | 1 | | 0.1874 | | 0.0525 | | 12.746 | 0.0004 |
| WOE\_PCT\_PROMISE\_1M | | 1 | | 0.3784 | | 0.0537 | | 49.5594 | <.0001 |
| WOE\_REDUCE\_RATE\_EAD\_C2 | | 1 | | 0.2773 | | 0.0548 | | 25.6252 | <.0001 |
| WOE\_CHECK\_REDUCE\_DPD\_C12 | | 1 | | 0.3184 | | 0.0466 | | 46.6063 | <.0001 |
| WOE\_COOPERATIONVPB\_TIME | | 1 | | 0.3822 | | 0.0789 | | 23.4342 | <.0001 |
| WOE\_REPAY\_AMT\_C3 | | 1 | | 0.3115 | | 0.0579 | | 28.923 | <.0001 |
| WOE\_CC\_INCREASE\_BAL\_C1 | | 1 | | 0.3184 | | 0.0461 | | 47.7206 | <.0001 |
| WOE\_PAYMENT\_3M | | 1 | | 0.2156 | | 0.0562 | | 14.7228 | 0.0001 |
|  | |  | |  | |  | |  |  |
|  | |  | |  | |  | |  |  |
| **Association of Predicted Probabilities and Observed** | | | | | | |
| **Responses** | | | | | | |
| Percent Concordant | 81.2 | | Somers' D | | 0.625 | |
| Percent Discordant | 18.7 | | Gamma | | 0.626 | |
| Percent Tied | 0.1 | | Tau-a | | 0.118 | |
| Pairs | 15367296 | | c | | 0.813 | |

### List of variables in model

|  |  |
| --- | --- |
| VARIABLE | DESCRIPTION |
| N\_DPD\_EOM\_CHANGE\_M3 | **Comparison customer's bucket at end of months T-3 with T-2, T-2 with T-1, and T-1 with T (T: observation date):**   * **Low: if bucket at end of T-3 >= bucket at end of T-2, similar for the couples (T-2, T-1), (T-1, T),** * **Up: if bucket at end of T-3 < bucket at end of T-2, similar for the couples (T-2, T-1), (T-1, T).** |
| CHECK\_REDUCE\_DPD\_C12 | **The number of months that customer reduced overdue during last 12 months before the observation date (only compare the DPD among end of months).** |
| N\_LOWDPD\_ALL\_C12 | **Number of times the customer had decreased overdue in last 12 months before observation date (Loan and Credit card)** |
| CC\_INCREASE\_BAL\_C1 | **The difference between total balance of customer’s credit card at observation date and total balance of customer’s credit card at 1 month before observation date** |
| PAYMENT\_3M | **Total customer’s repayment amount for credit card during 3 months before observation date** |
| REPAY\_AMT\_C3 | **Total amount paid to repay for loans during last 3 months before observation date** |
| REDUCE\_RATE\_EAD\_C2 | **The decrease of total balance of customer's loans in last 2 months divided by balance of all customer's loans at 2 months prior to the observation date** |
| PCT\_PROMISE\_1M | **The ratio between variable Number of times customer promised to pay during last month before observation date and number of actions (includes call, fields) during last month before observation date** |
| COOPERATIONVPB\_TIME | **Time with VPBank (Months)** |

### Explain variables in the model

N\_DPD\_EOM\_CHANGE\_M3:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| N\_DPD\_EOM\_CHANGE\_M3 | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| 01. LOW,LOW,LOW | 423 | 3.31% | 16.59% | 1.75% | 47.28% | 2.25 | 1.066 |
| 02. UP,LOW,LOW | 301 | 2.36% | 7.66% | 1.73% | 65.78% | 1.487 |
| 03. LOW,UP,LOW | 356 | 2.79% | 8.56% | 2.11% | 67.70% | 1.401 |
| 04. LOW,LOW,UP | 509 | 3.98% | 9.45% | 3.34% | 75.05% | 1.04 |
| 05. UP,UP,LOW | 704 | 5.51% | 11.09% | 4.85% | 78.84% | 0.826 |
| 06. UP,LOW,UP | 914 | 7.15% | 9.67% | 6.86% | 85.78% | 0.344 |
| 07. LOW,UP,UP | 1,609 | 12.59% | 12.65% | 12.59% | 89.43% | 0.005 |
| 08. UP,UP,UP | 7,962 | 62.31% | 24.33% | 66.77% | 95.89% | -1.01 |

Table : N\_DPD\_EOM\_CHANGE\_M3

The variable shows that the higher a customer jumps into the upper debt group, the worse it becomes.

CHECK\_REDUCE\_DPD\_C12:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CHECK\_REDUCE\_DPD\_C12 | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| 1. (Low; 0] | 8,259 | 64.63 | 29.61 | 68.75 | 95.18 | -0.842 | 0.842 |
| 2. (0; 1] | 2,381 | 18.63 | 22.32 | 18.2 | 87.4 | 0.204 |
| 3. (1; 2] | 1,104 | 8.64 | 18.9 | 7.43 | 76.99 | 0.933 |
| 4. (2; 3] | 605 | 4.73 | 15.1 | 3.52 | 66.45 | 1.458 |
| 5. (3; High) | 429 | 3.36 | 14.06 | 2.1 | 55.94 | 1.902 |

Table : CHECK\_REDUCE\_DPD\_C12

This variable shows the number of months that customer reduced overdue during last 12 months before the observation date. The more number of times customer lowering their overdue, the less probability of that customer go default.

N\_LOWDPD\_ALL\_C12

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| N\_LOWDPD\_ALL\_C12 | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| [1] LOW< - 5.6 | 4220 | 24.8 | 25.5 | 13.7 | 3.4 | 0.6 | 0.777 |
| [2] 5.6< - 6.9 | 2605 | 15.3 | 15.4 | 13.1 | 5.2 | 0.2 |
| [3] 6.9< - 7.7 | 1619 | 9.5 | 9.5 | 9.6 | 6.2 | 0.0 |
| [4] 7.7 < - HIGH | 8593 | 50.4 | 49.6 | 63.6 | 7.7 | -0.2 |

Table : Rloan\_Amt\_Income2

This is ratio the number of times the customer had decreased overdue in last 12 months before observation date (Loan and Credit card). The more number of times customer lowering their overdue, the less probability of that customer go default.

CC\_INCREASE\_BAL\_C1:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CC\_INCREASE\_BAL\_C1 | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| 01. No CC | 2,288 | 17.91% | 17.56% | 17.95% | 89.69% | -0.022 | 0.568 |
| 02. (Low; 0) | 1,110 | 8.69% | 29.54% | 6.24% | 64.23% | 1.555 |
| 03. [0; 200,000] | 775 | 6.07% | 10.57% | 5.54% | 81.68% | 0.646 |
| 04. (200,000; 500,000] | 908 | 7.11% | 7.51% | 7.06% | 88.88% | 0.063 |
| 05. (500,000; 2,500,000] | 6,203 | 48.54% | 29.09% | 50.83% | 93.70% | -0.558 |
| 06. (2,500,000; High) | 1,494 | 11.69% | 5.73% | 12.39% | 94.85% | -0.772 |

Table : CC\_INCREASE\_BAL\_C1

This variable is the difference between total balance of customer’s credit card at observation date and total balance of customer’s credit card at 1 month before observation date. The higher the difference, the higher probability of that customer go default.

PAYMENT\_3M:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PAYMENT\_3M | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| 01. No CC | 2,288 | 17.91% | 17.56% | 17.95% | 89.69% | -0.022 | 0.468 |
| 02. (Low; 150,000] | 5,728 | 44.83% | 21.13% | 47.61% | 95.04% | -0.812 |
| 03. (150,000; 500,000] | 541 | 4.23% | 2.75% | 4.41% | 93.16% | -0.471 |
| 04. (500,000; 1,500,000] | 1,142 | 8.94% | 9.60% | 8.86% | 88.70% | 0.08 |
| 05. (1,500,000; 2,000,000] | 401 | 3.14% | 4.17% | 3.02% | 86.03% | 0.323 |
| 06. (2,000,000; High) | 2,678 | 20.96% | 44.79% | 18.16% | 77.52% | 0.903 |

Table : PAYMENT\_3M

This variable is the total customer’s repayment amount for credit card during 3 months before observation date. The higher the amount, the higher probability of that customer go default.

REPAY\_AMT\_C3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| REPAY\_AMT\_C3 | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| 01. No loan/no payment | 7,428 | 58.13% | 48.29% | 59.29% | 91.26% | -0.205 | 0.432 |
| 02. (Low; 300,000] | 1,807 | 14.14% | 4.24% | 15.31% | 96.85% | -1.283 |
| 03. (300,000; 2,000,000] | 914 | 7.15% | 5.36% | 7.36% | 92.12% | -0.318 |
| 04. (2,000,000; 8,000,000] | 1,843 | 14.42% | 23.21% | 13.39% | 83.07% | 0.55 |
| 05. (8,000,000; 14,000,000] | 504 | 3.94% | 10.57% | 3.17% | 71.83% | 1.205 |
| 06. (14,000,000; High) | 282 | 2.21% | 8.33% | 1.49% | 60.28% | 1.724 |

Table : REPAY\_AMT\_C3

This variable is total amount paid to repay for loans during last 3 months before observation date. The higher amount was paid, the less probability of that customer go default.

REDUCE\_RATE\_EAD\_C2:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| REDUCE\_RATE\_EAD\_C2 | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| 1. New OD/Increase Bal | 4,862 | 38.05% | 25.07% | 39.57% | 93.07% | -0.456 | 0.376 |
| 2. No loan | 6,724 | 52.62% | 46.73% | 53.31% | 90.66% | -0.132 |
| 3. (-0.50; 3.5] | 550 | 4.30% | 10.64% | 3.56% | 74.00% | 1.095 |
| 4. (3.5; High) | 642 | 5.02% | 17.56% | 3.55% | 63.24% | 1.598 |

Table : REDUCE\_RATE\_EAD\_C2

This variable is the decrease of total balance of customer's loans in last 2 months divided by balance of all customer's loans at 2 months prior to the observation date. The higher the ratio, the less probability of that customer go default.

PCT\_PROMISE\_1M

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PCT\_PROMISE\_1M | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| 1. (Low; 0] | 9,602 | 75.14% | 52.75% | 77.78% | 92.62% | -0.388 | 0.33 |
| 2. (0; 0.1] | 558 | 4.37% | 5.88% | 4.19% | 85.84% | 0.339 |
| 3. (0.1; 0.3] | 1,414 | 11.07% | 18.08% | 10.24% | 82.81% | 0.568 |
| 4. (0.3; 0.5] | 791 | 6.19% | 12.95% | 5.40% | 78.00% | 0.875 |
| 5. (0.5; High)/ No call & field visit | 413 | 3.23% | 10.34% | 2.40% | 66.34% | 1.462 |

Table : PCT\_PROMISE\_1M

This variable is the ratio between variable Number of times customer promised to pay during last month before observation date and number of actions (includes call, fields) during last month before observation date. The higher the percentage, the less probability of that customer go default.

COOPERATIONVPB\_TIME:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| COOPERATIONVPB\_TIME | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| 01. (Low; 4] | 552 | 4.32% | 0.52% | 4.77% | 98.73% | -2.214 | 0.304 |
| 02. (4; 5] | 596 | 4.66% | 0.82% | 5.12% | 98.15% | -1.833 |
| 03. (5; 9] | 1,455 | 11.39% | 5.28% | 12.10% | 95.12% | -0.829 |
| 04. (9; 20] | 2,318 | 18.14% | 15.18% | 18.49% | 91.20% | -0.197 |
| 05. (20; 35] | 2,092 | 16.37% | 17.86% | 16.20% | 88.53% | 0.098 |
| 06. (35; 100] | 5,101 | 39.92% | 50.97% | 38.62% | 86.57% | 0.277 |
| 07. (100; High) | 664 | 5.20% | 9.38% | 4.71% | 81.02% | 0.689 |

Table : COOPERATIONVPB\_TIME

This variable is the customer’s time with VPBank. The longer a customer stays, the less probability of that customer go default.

## Model statistic assessment

### Model discrimination

Model performance:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Score band | Total, # | Good, # | Bad, # | Bads% in band | %Bads in band & over |
| B1 | 724 | 717 | 7 | 99.0 | 89.5 |
| B2 | 576 | 567 | 9 | 98.4 | 88.9 |
| B3 | 1280 | 1254 | 26 | 98.0 | 88.4 |
| B4 | 1359 | 1321 | 38 | 97.2 | 87.2 |
| B5 | 1293 | 1253 | 40 | 96.9 | 85.7 |
| B6 | 1852 | 1758 | 94 | 94.9 | 83.8 |
| B7 | 1227 | 1146 | 81 | 93.4 | 80.2 |
| B8 | 645 | 593 | 52 | 91.9 | 76.5 |
| B9 | 656 | 576 | 80 | 87.8 | 73.9 |
| B10 | 1252 | 1042 | 210 | 83.2 | 71.0 |
| B11 | 643 | 482 | 161 | 75.0 | 63.1 |
| B12 | 642 | 411 | 231 | 64.0 | 57.0 |
| B13 | 629 | 314 | 315 | 49.9 | 49.9 |

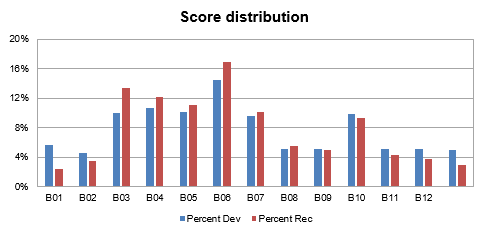
Table : Score band

Table : Score band

### Model accuracy

# Pre-validation

## Model stability



# Model makers

List of staff involved in model development

|  |  |  |
| --- | --- | --- |
| **Task** | **Main PIC** | **Support** |
| Nguyễn Thị Hạ Hương | huongnth1@vpbank.com.vn | Project management |
| Trần Quốc Thái | thaitq15@vpbank.com.vn | Model owner |
| Trần Quốc Thái | thaitq15@vpbank.com.vn | Extract variable/Data quality |
| Trần Quốc Thái | thaitq15@vpbank.com.vn | Model Validator |
| Trần Quốc Thái | thaitq15@vpbank.com.vn | Test UAT |

Table 18: List of participants

# Appendix

# Bibliography

**There are no sources in the current document.**