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

**APPLICATION SCORECARD FOR UPL**

**Model development report**

Modelling Team

December, 2019

|  |  |
| --- | --- |
| Author: |  |
| Contact: |  |
| Creation Date: | 2019/12/24 |
| Last Modification Date: |  |
| Version: | 1.5 |
| Status: | Final Version |

**Document Control**

**Change Record**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Author** | **Version** | **Change Reference** |
| 2018/10/22 |  | 1.0 | Initial Version |
| 2019/08/07 |  | 1.4 | Update contents |
| 2019/12/20 |  | 1.5 | Update contents |
|  |  |  |  |

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 July 2017 to June 2018 | | |
| Title | Corporate borrower rating model | | |
| Version | 1.0 (First generation) | | |
| Portfolio for model | Corporate | | |
| List of systems on which the model is implemented | The model was implemented in the Risk Rating Tool (RRT) on 08/15 of February 2020.  <Mô tả chức năng rating của hệ thống:   * Đối với A-score: mô tả hệ thống implement mô hình * Đối với các mô hình khác chạy thủ công trên pc hoặc tự động qua job sql, sas thì mô tả sơ lược quy trình chấm điểm này>   Risk Rating Tool (RRT):   * keeps information on financial statements and qualitative information of borrowers; * determines borrower's rating (score); * …   Core banking system – calculates Delinquency. The system is managed by IT department; | | |
| Date of implementation | <ngày implement trên hệ thống/ngày bắt đầu sử dụng mô hình>  Feb 8th 2020 | | |
| Brief description of the model | <Mô tả thông tin sơ lược về mô hình/hệ thống xếp hạng đã được xây dựng, bao gồm:   * Quy trình xếp hạng * Phương pháp sử dụng để xây dựng mô hình * Độ dài của mẫu dev và outcome period>   The model consists of two modules:   1. Quantitative module (9 factors) 2. Qualitative module (5 factors)   Before assigning a rating to the borrower, the model checks it against “knock out” factors. If they are triggered, the default rating is automatically attributed to the borrower.  Prior to assigning the "final rating" to the borrower, the "first rating" can be adjusted by applying “Adjustment/Override factors” if any of specified for them criteria were met.  Logit regression method was used in the model for predicting the creditworthiness of the borrowers.  The model includes observation period which is less than 5 years due to the lack of data. The used outcome period depending on the customers is equal or less than one year. | | |
| 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 | <đơn vị sử dụng mô hình>  CBR, CIB-CMB | | |
| Approved by | Scoring committee (CRO, CEO) | | |
| Date of approval | July 31st  2023 | | |
| Head of department responsible for the model | Minh Chu Hong (RMD - CrMD) <minhch@vpbank.com.vn> | | |
| Employee responsible for the model | <người chịu trách nhiệm xây dựng mô hình> | | |
| Portfolio statistics **as of 31.12.2019 <thời điểm xây dựng mô hình>** | |  | |
| Number of clients | 438 | | |
| Loan portfolio amount | 117,255,567,979,393 VND | | |
| Development data statistics (mẫu xây dựng mô hình) | |  | |
| Development sample | 01.01.2016 – 31.12.2016 | | |
| Good / bad / total observations | 152 / 36 / 188 | | |
| Gini on development sample | 70% | | |
| In time validation data statistics (mẫu 30% - nếu có) | | | |
| In time validation sample | 31.12.2016-31.12.2018 | | |
| Good / bad / total observations | 1026 / 23 / 1049 | | |  |
| Gini on development sample | 5% | | |  |
| Initial validation data statistics (mẫu OOT) | | | |
| Validation sample | 31.12.2016-31.12.2018 | | |
| Good / bad / total observations | 1026 / 23 / 1049 | |  |
| Gini on validation sample | 5% | |  |
| Stability analysis | | | |
| Recent sample | 31.12.2016-31.12.2019 | |  |
| PSI on recent sample | 5% | |  |

Table 1: Executive summary

# Model design & Model scope

## Model design

<với A-score của Retail, mô hình chỉ là score card thông thường, với A-score của SME thì bao gồm KO, adjustment sau khi mô hình chấm xong, đối với B-score cũng là score card thông thường. Đối với mô hình không cần mô tả phần design thì để toàn bộ mục 2 là methodology>

## 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

<Mô tả về phương pháp xây dựng mô hình, đối với mô hình scorecard thông thường thì có các phần sau: phân tích đơn biến sử dụng WOE, phân tích tương quan, phân tích đa biến, logistic regression, các tiêu chuẩn để đánh giá mô hình. **Chỉ trình bày trong document những phần mà** **mình dùng để develop**, **không trình bày thừa/thiếu**. >

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

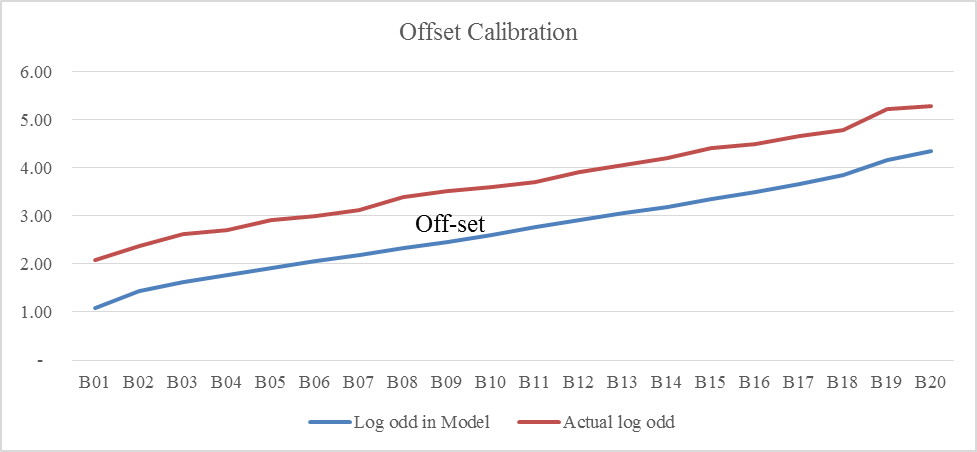
Below does Modelling Team recommend the benchmark of PSI:

* : 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

<Yêu cầu của phần này là tính minh bạch, kế thừa, tức là những gì viết trong document cần đủ để người khác có thể thực hiện lại việc xây dựng mô hình.>

## Target definition

### Detail definition of target

The Good/Bad/Indeterminate is defined as following:

* Good observation: never 30+ DPD in outcome period
* Bad observations: 60+ DPD during outcome period.
* Indeterminate observations: others

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

|  |  |  |
| --- | --- | --- |
|  | **Frequency** | **Percent** |
| Bad: 60+ DPD ever | 1 111 | 6.2% |
| Good: never 30+ DPD | 16 073 | 90.1% |
| Indeterminate: Other | 652 | 3.7% |
| **TOTAL** | **17 836** | **100%** |

Table 5: Training sample overview

Performance window was chosen with 9 months from the disbursement date.

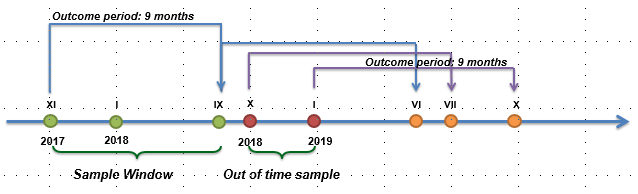


Figure 2: Observation window and performance window

### Confirm the target definition

<Giải thích tại sao lại sử dụng định nghĩa này>

<Đối với các mô hình sử dụng với yêu cầu tuân thủ cần nêu rõ định nghĩa này có đủ để tuân thủ hay không?>

#### Roll rate analysis

Roll rate analysis was made in order to find appropriate definition of Good and Bad. The report below is done on sample of New to bank unsecured loans existing at end of June 2019. The bucket in vertical axis shows the maximum DPD of loans during last 1 year. The bucket in horizotal axis shows show the maxium DPD of loans during next 1 year.

Figure 3: Roll rate analysis

The Roll rate analysis shows that:

* Bucket 0 DPD: 87.6% loans stay in the same bucket or move to bucket 0 – 10 DPD, 3.7% loans become 90+ after one year.
* Bucket 0 – 10 DPD: 73.2% loans stay in the same bucket or move to lower bucket, 7% loans become 90+ DPD.
* Bucket 10 – 30 DPD: 56.9% loans stay in the same bucket or move to lower bucket, 24.6% loans become 90+ DPD.
* Bucket 30 – 60 DPD: 53.4% loans move to higher DPD bucket, 15.3% loans stay in the same bucket, 31.2% loans move back to lower DPD bucket.
* Bucket 60 - 90 DPD: 69.8% loans move to higher DPD bucket, 12.3% loans stay in the same bucket, 17.9% loans move back to lower DPD bucket.

Therefore, the Good/Bad/Indeterminate is defined as following:

* Good observation: never 30+ DPD in outcome period
* Bad observations: 60+ DPD during outcome period.
* Indeterminate observation: other

#### Vintage analysis

Vintage 90+ was made in order to find outcome period of model. The report was done on all New to bank unsecured loans disbursed from November 2017 to April 2019.

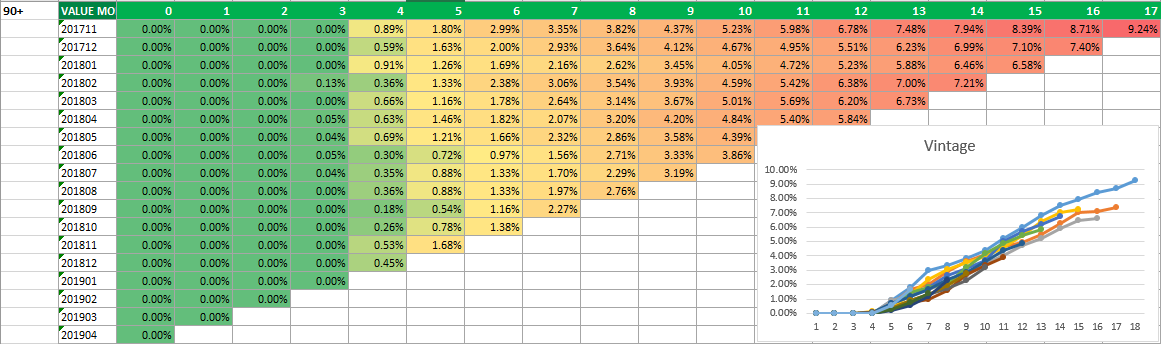


Figure 4: Vintage analysis

## Data preparation

### Data source

<Mô tả sơ lược về các nguồn thông tin được lấy để xây dựng mô hình>

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

* Personal information: is information about the demographics of borrowers. These information are in loan application form and other documents provided by customers. These information are inputted in Finnone system or RLOS by CRS and consolidated in DWH by IT. Modelling team queries data directly from DWH.
* Behavioral information: is information about behavior of borrowers including information of loans, deposits, current accounts, credit cards and repayment. It is information about the credit and non-credit relationship of customers with VPBank before application date. These data are taken from the source tables on DWH and calculated at the time of applying for a loan.
* 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.

Details of data sources are shown in the table below:

<Chi tiết các bảng dùng để lấy dữ liệu xây dựng mô hình>

|  |  |
| --- | --- |
| Group - variable | Source |
| Demographic variables | [10.36.13.74].BICDATA.dbo.FINNONE\_LIVE |
| [10.36.13.74].STAGING.dbo.F1TB\_NBFC\_CUSTOMER |
| [10.36.13.74].STAGING.dbo.F1TB\_NBFC\_CUSTOMER\_SAL\_HEAD |
| [10.36.13.74].STAGING.dbo.F1TB\_NBFC\_ADDRESS |
| [10.36.31.16].VPB\_WHR2.dbo.TBL\_RLOS\_APP |
| [10.36.31.16].VPB\_WHR2.dbo.TBL\_RLOS\_APP\_FORM |
| Loan | [10.36.31.16].VPB\_WHR2.dbo.T24CRD |
| [10.36.31.16].VPB\_WHR2.dbo.T24CRD\_DISBTDY\_ALL |
| [10.36.31.16].VPB\_WHR2.dbo.T24CRD\_LD\_SCHEDULE\_NEW |
| [10.36.13.74].BICDATA.dbo.LNTB\_MASTER\_ENDMONTH |
| Credit card | [10.36.13.74].BICDATA.dbo.CARD\_ENDMONTH |
| [10.36.31.16].VPB\_WHR2.dbo.TBL\_W4\_CONTRACT |
| Deposit & Current account | [10.36.13.74].BICDATA. dbo.DPTB\_MASTER\_ENDMONTH |
| [10.36.31.16].VPB\_WHR2.dbo.FOCURR\_SAVE |
| [10.36.31.16].VPB\_WHR2.dbo.FT\_HIST\_2019 |
| [10.36.31.16].VPB\_WHR2.dbo.TT\_HIST |
| PCB information | [10.36.28.85].VPB\_CIC\_INTERNAL.dbo.[T\_PCB\_FINAL\_RESUL] |
| [10.36.31.16].VPB\_WHR2.dbo.TBL\_RLOS\_APP\_PCB\_WIDEGT |
| File excel from PCB |
| Other | [10.36.31.16].VPB\_WHR2. dbo.VPB\_CUSTOMER |

Table 2: Data source

### Sample selection

<thời gian lấy mẫu xây dựng mô hình, phân tách rõ mẫu dev và mẫu OOT>

The model data includes 17836 loan applications disbursed from November 2017 to September 2018 in the New to Bank segment. After data filtering, the data in the model was composed of 17037 applications. Details of the filter process will be presented later.

|  |  |  |
| --- | --- | --- |
| **Data** | **Time frame** | **#Observations** |
| Development | 11.2017-09.2018 | 17037 |
| Out of time validation | 10.2018-01.2019 | 8605 |

Table 3: Sample overview

<Giải thích tại sao lại chọn mẫu này và đánh giá mẫu này có tính đại diện hay không.>

<Với các mô hình sử dụng với mục đích tuân thủ, cần phải đảm bảo mẫu này có thỏa mãn tính tuân thủ hay không?>

Bad rate analysis

In order to determinate sample window, we made analysis of bad rate over than two years history data. (Bad is considered 30+ DPD at the end of next 3 months from disbursement date). The selection period satisfy:

* No effect of seasonality
* To be close to development time

Figure 1: Bad rate analysis

As can be seen from above chart, bad rate has no significant change from November 2017. 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 9 months from the time of building the model. Therefore, the period from November 2017 to September 2018 is proposal sample window.

### Data exclusion

<Trình bày các bước lọc để đưa ra mẫu xây dựng mô hình/mẫu validate cuối cùng. Tại mỗi bước nêu lý do tại sao lại lọc các quan sát này.>

Data filter is performed in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Sample** | **# Observations** | **%Fitter** |
| 1 | UPL loans except Household business, disbursed from Nov 1st 2017 to Sep 30th 2018 | 17836 | 100% |
| 2 | Remove observations which belong to Bad bank list | 17802 | 99.8% |
| 3 | Remove observations which belong to fraud list | 17676 | 99.1% |
| 4 | Remove indeterminate | 17037 | 95.5% |

Table 4: Data filter

Step 1: All loans disbursed from 11.2017 to 09.2018 in the New to Bank segment, except household business, was taken as sample of model

Step 2: Remove observations which belong to Bad bank list

Step 3: Remove observations which belong to fraud list

Step 4: Remove indeterminate

### Variable description

<Danh sách các biến thu thập để xây dựng mô hình chia theo từng mảng dữ liệu, mô tả tính đầy đủ của các biến này về: các mảng dữ liệu liên quan đến chất lượng tín dụng của KH, về khoảng thời gian quan sát (1-12 tháng). Giải thích các trường hợp không lấy được mảng biến nào đó nếu có>

### Data quality

<Đánh giá chất lượng dữ liệu dùng cho việc xây dựng mô hình, bao gồm dữ liệu cho target và dữ liệu cho các biến giải thích>

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

## Segmentation

<Lý do thực hiện segmentation, cách thức thực hiện và kết quả của việc segment, mô tả các mẫu nhận được bao gồm số quan sát, bad rate, các đặc trưng khác của mỗi mẫu, hiệu quả của việc dùng segmentation>

Both experience-based and statistical analyses will yield ideas for potential segmentation, and may confirm that there are sufficient reasons for segmenting—but they do not quantify the benefits of building multiple segmented models. There are fairly simple ways available to estimate whether the improvement through segmentation is worth pursuing.

The following exercise involves building “quick and dirty” models for a base case/unsegmented data set, as well as for the proposed segments, and then comparing their performances. The numbers generated are therefore ballpark estimates that can be used for comparison purposes without spending months building more fine-tuned scorecards.

The first step is to measure the improvement in predictive power through segmentation, as one of our reasons for building segmented models was the belief that segmentation maximizes predictiveness for unique segments. This can be done using a number of statistics such as the KS, c-statistic and so on.

[Table: Comparing improvement]

## Single factor analysis

<phần này mô tả kết quả của việc phân tích đơn biến trong mô hình này, các đặc điểm riêng của mô hình, bao gồm: xử lý missing, loại biến không phân tích, đánh giá từng biến có phù hợp hay không phù hợp với business, đặc trưng riêng của biến. Tập hợp lại theo bảng để dễ theo dõi, chi tiết trong file excel đính kèm. Đặc biệt chú ý các thông tin được sử dụng để nhóm biến ngoài việc phân nhóm theo rule của WOE. Cần hiển thị bao nhiêu biến được phân tích và bao nhiêu biến phù hợp để đưa vào phân tích đa biến. >

<Đối với mỗi threshold không phổ biến cần có các phân tích để giải thích lý do sử dụng threshold này>

## Multi factor analysis

<trong phần này mô tả trực tiếp các bước để đưa ra long list variable dùng cho logistic regression + giải thích>

## Logistic regression

<Trình bày cách thức thực hiện dẫn đến mô hình cuối cùng + các tham số đưa vào regression và + kết quả regression + giải thích ý nghĩa của từng biến trong mô hình. Hiện tại team có nhiều cách thức để đưa ra mô hình cuối cùng, ví dụ:

* Chạy thủ công: cần nêu các bước chạy ra mô hình
* Chạy theo marginal: cần nêu rõ kq dùng marginal, lấy ds biến từ kq marginal để chạy lại theo logistic stepwise

Các bước loại biến cần nêu rõ lý do. >

### Result of logistic regression in SAS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analysis of Maximum Likelihood Estimates | | | | | |
| Parameter | **DF** | **Estimate** | **Standard** | **Wald** | **Pr > ChiSq** |
| **Error** | **Chi-Square** |
| Intercept | 1 | 2.7413 | 0.0352 | 6058.1512 | <.0001 |
| WOE\_AGE\_IDCARD | 1 | 0.8822 | 0.1516 | 33.8512 | <.0001 |
| WOE\_RLOAN\_AMT\_INCOME2 | 1 | 1.0928 | 0.1073 | 103.7944 | <.0001 |
| WOE\_TIME\_WITH\_VPB | 1 | 0.5039 | 0.2075 | 5.8997 | 0.0151 |
| WOE\_BALANCE\_PCT\_LIMIT\_CC\_ALL | 1 | 0.7809 | 0.1425 | 30.0357 | <.0001 |
| WOE\_USE\_SERVICE\_CUR\_ALL | 1 | 0.7495 | 0.1506 | 24.7658 | <.0001 |
| WOE\_RISK\_ZONE2 | 1 | 1.2012 | 0.1952 | 37.8577 | <.0001 |
| WOE\_PD | 1 | 0.5165 | 0.185 | 7.7919 | 0.0052 |
| WOE\_CR\_WORST\_36M\_TIME\_WORST | 1 | 0.9725 | 0.0772 | 158.6869 | <.0001 |
| WOE\_CR\_MAXDPD\_1Y\_MOB | 1 | 0.6788 | 0.2036 | 11.1148 | 0.0009 |
| WOE\_CR\_GENDER\_MARITAL | 1 | 1.0318 | 0.111 | 86.4732 | <.0001 |
| WOE\_CR\_AGE\_OCCUPATION | 1 | 0.9446 | 0.102 | 85.8297 | <.0001 |
| WOE\_CR\_AGE\_MARITAL | 1 | 0.4751 | 0.1758 | 7.3073 | 0.0069 |

|  |  |  |  |
| --- | --- | --- | --- |
| Association of Predicted Probabilities and Observed | | | |
| Responses | | | |
| Percent Concordant | 73 | **Somers' D** | 0.46 |
| Percent Discordant | 27 | **Gamma** | 0.46 |
| Percent Tied | 0.1 | **Tau-a** | 0.053 |
| Pairs | 16621922 | **c** | 0.73 |

### List of variables in model

|  |  |
| --- | --- |
| VARIABLE | DESCRIPTION |
| Cross1: Cr\_Worst\_36m\_Time\_Worst | Cross variable between Worst\_Status\_36m and Time\_Worst\_Stts\_All\_M - Worst\_Status\_36m: Worst status within last 36 months - Time\_Worst\_Stts\_All\_M: Number of month from creation date back to the time when customer was in highest debt group (for all product) in the past |
| Cross2: Cr\_Age\_Occupation | Cross variable between Age and Occupation |
| Rloan\_Amt\_Income2 | Ratio of request loan amount and total outstanding loans of customer at VPBank and other Institutions divided by customer‘s monthly income |
| Cross3: Cr\_Gender\_Marital | Cross variable between Gender and Marital\_Status |
| Cross4: Cr\_Maxdpd\_1y\_Mob | Cross variable between Maxdpd\_1y\_Bfrappl and Mob - Maxdpd\_1y\_Bfrappl: Maximum delinquency of customer for all his loans during last 1 year before application date - Mob: number of month on book of latest loan in VPBank |
| Balance\_Pct\_Limit\_Cc\_All | Total balance of all credit cards of customers at VPBank and other Institutions divided by total of their limits |
| Use\_Service\_Cur\_All | Type of service that customer currently uses at VPBank and other Institutions as of creation date, if customers has multi product, use priority from Depo -> Payroll -> Secured -> Unsecured -> Credit card -> Current account |
| Pd | Probability of default of customer which is generated from in-house behavior models (not include credit card behavior model) at end of last month before creation date |
| Cross5: Cr\_Age\_Marital | Cross variable between Age and Marital Status |
| Age\_Idcard | Age of customer's identification  Age\_idcard = round((creation\_date - date\_of\_issue)/365,5) |
| Time\_With\_Vpb | Number of month from CIF open date at VPBank to creation date |
| Risk\_Zone2 | Risk level of branches/BU at VPBank according to OVD 90+ ratio calculated for all loans which were disbursed during last 6 months. |

### Explain variables in the model

<Đối với mỗi biến cần trình bày:

Table: binning, number of good, number of bad, WOE, Bad rate, IV

Hình minh họa: Number of good, number of bad, WOE

Giải thích: cách nhóm biến, tính phù hợp với business>

Cross1: Cr\_Worst\_36m\_Time\_Worst:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CR\_WORST\_36M\_TIME\_WORST | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| GROUP 1 | 3322 | 19.5 | 20.0 | 12.3 | 3.9 | 0.5 | 0.15 |
| GROUP 2 | 5175 | 30.4 | 30.5 | 28.8 | 5.8 | 0.1 |
| GROUP 3 | 917 | 5.4 | 5.0 | 11.5 | 13.0 | -0.8 |
| GROUP 4 | 249 | 1.5 | 1.2 | 5.2 | 21.7 | -1.5 |
| NO\_PCB | 7374 | 43.3 | 43.3 | 42.3 | 6.0 | 0.0 |

Table 6: Cr\_Worst\_36m\_Time\_Worst

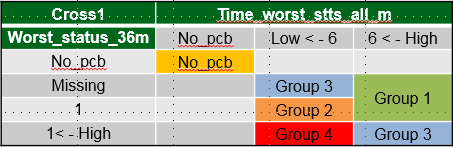


Figure 8: Cr\_Worst\_36m\_Time\_Worst

This variable is the cross variable between the highest debt group of customers in last 36 months credit relationship history at PCB and number of months from the time when customer had the highest debt group to application date.

The "No\_pcb" group is customers group with no PCB information on the debt group, which do not have any basis to assess the repayment capacity. This group has a bad rate of 6.0%, close to the average bad rate of the sample. Customers in this group have no increase or decrease in score.

The “Group 4” group is customers who have been in greater or equal 2 debt group (day past due > 9) during last 6 month before application date. This group is the worst group with a very high bad rate of 21.7%, which is nearly 3.5 times as high as average bad rate of the whole sample.

Group "Group 3" is the group of customers who have been in greater 1 debt group during a period of greater than 6 months or a new group of customers apply loan or credit card at other institution and is rejected. This group has a bad rate of 13.0%, which is more 2 times as high as average bad rate of the whole sample.

Other groups have a gradual decrease in bad rate corresponding to the decrease in debt group and the increase in number of months from the time when customer had the highest debt group to application date.

Cross2: Cr\_Age\_Occupation:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CR\_AGE\_OCCUPATION | # G+B |  | % | % Goods | % Bads | Bad rate | WOE | IV |
| GROUP 1 | 1939 |  | 11.4 | 10.8 | 19.8 | 10.6 | -0.6 | 0.12 |
| GROUP 2 | 2161 |  | 12.7 | 12.5 | 15.9 | 7.6 | -0.2 |
| GROUP 3 | 9462 |  | 55.5 | 55.8 | 51.5 | 5.7 | 0.1 |
| GROUP 4 | 2994 |  | 17.6 | 17.9 | 12.1 | 4.2 | 0.4 |
| GROUP 5 | 481 |  | 2.8 | 3.0 | 0.7 | 1.5 | 1.5 |

Table 7: Cr\_Age\_Occupation

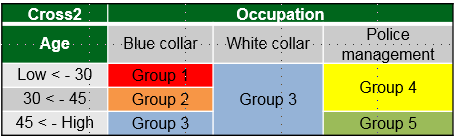


Figure 9: Cr\_Age\_Occupation

The customers become older, financial capability is more stable so the ability to repay is also better. The customers who work in better positions, have higher incomes and better repayment capacity. Data is grouped according to this logic.

Rloan\_Amt\_Income2:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| RLOAN\_AMT\_INCOME2 | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| [1] LOW< - 5.6 | 4220 | 24.8 | 25.5 | 13.7 | 3.4 | 0.6 | 0.11 |
| [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 8: Rloan\_Amt\_Income2

This is ratio of request loan amount and total outstanding loans of customer at VPBank and other Institutions divided by customer‘s monthly income. This ratio represents the customer's liability repayment, the higher the ratio, the lower the repayment capacity.

Cross3: Cr\_Gender\_Marital:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CR\_AGE\_MARITAL | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| GROUP 1 | 1763 | 10.3 | 10.1 | 13.6 | 8.0 | -0.3 | 0.05 |
| GROUP 2 | 6942 | 40.7 | 40.6 | 43.4 | 6.5 | -0.1 |
| GROUP 3 | 6798 | 39.9 | 40.0 | 38.4 | 5.9 | 0.0 |
| GROUP 4 | 1534 | 9.0 | 9.3 | 4.6 | 3.1 | 0.7 |

Table 9: Cr\_Gender\_Marital

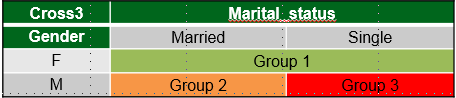


Figure 10: Cr\_Gender\_Marital

This variable is cross variable between gender and marital status of customer. Female customers have repayment behavior better than male customers. Married customers have repayment ability better than single customers.

Cross4: Cr\_maxdpd\_1y\_mob:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CR\_MAXDPD\_1Y\_MOB | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| MISSING | 14964 | 87.8 | 87.6 | 90.8 | 6.3 | 0.0 | 0.07 |
| GROUP 2 | 1267 | 7.4 | 7.4 | 8.4 | 6.9 | -0.1 |
| GROUP 1 | 806 | 4.7 | 5.0 | 0.9 | 1.1 | 1.8 |

Table 10: Cr\_maxdpd\_1y\_mob

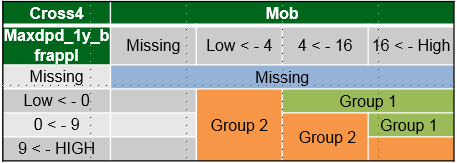


Figure 11 Cr\_maxdpd\_1y\_mob

This variable is cross variable between maximum delinquency of customer for all loans of customer during last 1 year before application date and number of month on book of latest loan of customer in VPBank.

The “Missing” group is customer group had no loan during last 1 year before application date. This group has a bad rate of 6.3%, which is close to the average bad rate of the whole sample.

Balance\_Pct\_Limit\_Cc\_All:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| BALANCE\_PCT\_LIMIT\_CC\_ALL | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| [1] MISSING | 14103 | 82.8 | 82.5 | 86.9 | 6.4 | -0.1 | 0.07 |
| [2] LOW < - 0.8 | 2029 | 11.9 | 12.3 | 6.1 | 3.1 | 0.7 |
| [3] 0.8 < - 0.96 | 454 | 2.7 | 2.7 | 1.9 | 4.4 | 0.3 |
| [4] 0.96 < - HIGH | 451 | 2.6 | 2.5 | 5.1 | 11.8 | -0.7 |

Table 11: Balance\_Pct\_Limit\_cc\_all

This variable is ratio of total balance of all credit cards of customers at VPBank and other Institutions divided by total of their limits. Xu hướng của biến chỉ ra rằng khách hàng có tỷ lệ chi tiêu trên thẻ càng cao thì có khả năng trả nợ thấp hơn. Điều này phù hợp với thực tế. Những khách hàng có tỷ lệ chi tiêu thẻ lên đến trên 96% hạn mức thể hiện khách hàng đang có nhu cầu sử dụng tiền rất cao và có thể đang gặp khó khăn về vấn đề tài chính.

Use\_Service\_Cur\_All:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| USE\_SERVICE\_CUR\_ALL | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| MISSING | 7130 | 41.9 | 41.7 | 43.4 | 6.3 | 0.0 | 0.06 |
| CREDITCARD\_CURRENT\_ACC | 3528 | 20.7 | 21.0 | 16.2 | 4.8 | 0.3 |
| DEPO\_PAYROLL\_SECURED | 930 | 5.5 | 5.7 | 2.2 | 2.5 | 0.9 |
| UNSECURED | 5449 | 32.0 | 31.6 | 38.2 | 7.3 | -0.2 |

Table 12: Use\_Service\_Cur\_All

This variable is type of service that customer currently uses at VPBank and other Institutions as of application date.

The “Depo\_payroll\_secured” group is the best group with a bad rate of 2.5%, approximately 1/3 of the average bad rate of the whole sample. This is customers group have at least one deposit or one payroll account at VPBank or a secured loan.

The “Unsecured” group is customers group have at least one unsecured loan. Interest rates of unsecured loans are relatively high. In fact, if it is not necessary, when the customer is having an unsecured loan, the customer will not borrow another unsecured loan. Therefore, this group is the worst group with a bad rate of 7.3% is reasonable.

The “Missing” group is customers group is not having credit relationship at VPBank and other institution. This group has a bad rate of 6.3%, close to the average bad rate of the sample. Customers in this group have no increase or decrease in score.

Pd:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PD | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| [1] MISSING | 15719 | 92.3 | 92.1 | 95.3 | 6.3 | 0.0 | 0.05 |
| [2] LOW < - 0.098 | 1057 | 6.2 | 6.5 | 2.2 | 2.2 | 1.1 |
| [3] 0.098 < - HIGH | 261 | 1.5 | 1.5 | 2.5 | 10.0 | -0.5 |

Table 13: PD

This variable is probability of default of customer which is generated from in-house behavior models (not include credit card behavior model) at end of last month before creation date. The greater the customer's PD, the lower the repayment capacity.

Cross5: Cr\_Age\_Marital:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CR\_AGE\_MARITAL | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| GROUP 1 | 1763 | 10.3 | 10.1 | 13.6 | 8.0 | -0.3 | 0.05 |
| GROUP 2 | 6942 | 40.7 | 40.6 | 43.4 | 6.5 | -0.1 |
| GROUP 3 | 6798 | 39.9 | 40.0 | 38.4 | 5.9 | 0.0 |
| GROUP 4 | 1534 | 9.0 | 9.3 | 4.6 | 3.1 | 0.7 |

Table 14: Cr\_age\_marital

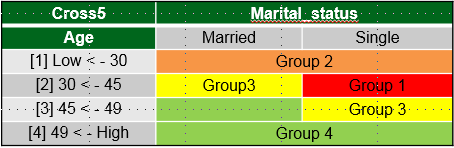


Figure 12: Cr\_age\_marital

Age\_Idcard:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AGE\_IDCARD | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| [1] LOW < - 0.75 | 1734 | 10.2 | 9.9 | 15.2 | 9.1 | -0.4 | 0.04 |
| [2] 0.75 < - 8.5 | 11012 | 64.6 | 64.5 | 66.0 | 6.2 | 0.0 |
| [3] 8.5 < - HIGH | 4291 | 25.2 | 25.6 | 18.8 | 4.5 | 0.3 |

Table 15: Age\_idcard

The table above shows that the higher time from issue data of ID card is, the better customer is.

Reason: Citizens with aged>= 16 years old have been ID card. Customers with bad credit histories want to hide this situation, they can remake ID card to apply for a loan to another bank

Time\_with\_vpb:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TIME\_WITH\_VPB | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| [1] NO RELATIONSHIP | 10347 | 60.7 | 60.5 | 64.9 | 6.5 | -0.1 | 0.03 |
| [2] 0.07 < - 9 | 2442 | 14.3 | 14.2 | 16.0 | 6.8 | -0.1 |
| [3] 9 < - 24.5 | 1696 | 10.0 | 10.0 | 9.4 | 5.8 | 0.1 |
| [4] 24.5 < - HIGH | 2552 | 15.0 | 15.3 | 9.7 | 4.0 | 0.5 |

Table 16: Time\_with\_vpb

Time\_cus\_open: The longer time you have a relationship with VPBank, the better your ability to repay you have. For the group “No relationship" Customers have never had a credit relationship with VPBank. These customers are new customers, and we have not had a basis for assessing the repayment capacity of our customers, therefore, it is reasonale to put this group of customers with bad rate of 6.5%, close to the average bad rate of the development sample. Furthermore, seeking new customers is one of the basic tasks of business, we can not deduct too much score for these customers.

Risk\_zone2:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| RISK\_ZONE2 | # G+B | % | % Goods | % Bads | Bad rate | WOE | IV |
| GROUP 1 | 12183 | 71.5 | 72.0 | 64.3 | 5.5 | 0.1 | 0.03 |
| GROUP 2 | 4854 | 28.5 | 28.0 | 35.7 | 7.6 | -0.2 |

Table 17: Risk\_zone2

## Calibration/transformation

<- Mô tả phương thức dùng để calibrate mô hình hoặc transform thành điểm, đối với mô hình application

* Kết quả của việc calibration: master scale và PD tương ứng, phân phối trên mẫu dev
* Kết quả của việc biến đổi thành điểm: score cho từng attribute.>

Ví dụ đối với việc biến đổi thành điểm:

In order to implement model in system, model need to be transformed into score with assumption followings:

1. Score is 500 for odd = 2.25
2. Score increase 100 when double odd.

Score calculation formula after calibration:

**Alpha** = 393.898

**Beta** = 141.092

**Score** = round ((alpha + good2 \* beta), 0)

Where: Good2 = log (odds)

Detail how to calculate the alpha, beta coefficient in file “Calibration\_upl\_pcb\_2018.xlsx”.

Scores for each attribute of each variable are shown below:

|  |  |
| --- | --- |
| CR\_WORST\_36M\_TIME\_WORST | score\_for\_attribute |
| GROUP 1 | 265 |
| GROUP 2 | 207 |
| GROUP 3 | 85 |
| GROUP 4 | 0 |
| NO\_PCB | 203 |

|  |  |
| --- | --- |
| CR\_AGE\_OCCUPATION | score\_for\_attribute |
| GROUP 1 | 0 |
| GROUP 2 | 48 |
| GROUP 3 | 91 |
| GROUP 4 | 132 |
| GROUP 5 | 277 |

|  |  |
| --- | --- |
| RLOAN\_AMT\_INCOME2 | score\_for\_attribute |
| [1] LOW< - 5.6 | 134 |
| [2] 5.6< - 6.9 | 63 |
| [3] 6.9< - 7.7 | 36 |
| [4] 7.7 < - HIGH | 0 |

|  |  |
| --- | --- |
| CR\_GENDER\_MARITAL | score\_for\_attribute |
| GROUP 1 | 110 |
| GROUP 2 | 31 |
| GROUP 3 | 0 |

|  |  |
| --- | --- |
| CR\_MAXDPD\_1Y\_MOB | score\_for\_attribute |
| GROUP 1 | 180 |
| GROUP 2 | 0 |
| MISSING | 9 |

|  |  |
| --- | --- |
| BALANCE\_PCT\_LIMIT\_CC\_ALL | score\_for\_attribute |
| [1] MISSING | 73 |
| [2] LOW < - 0.8 | 157 |
| [3] 0.8 < - 0.96 | 117 |
| [4] 0.96 < - HIGH | 0 |

|  |  |
| --- | --- |
| USE\_SERVICE\_CUR\_ALL | score\_for\_attribute |
| CREDITCARD\_CURRENT\_ACC | 48 |
| DEPO\_PAYROLL\_SECURED | 119 |
| MISSING | 16 |
| UNSECURED | 0 |

|  |  |
| --- | --- |
| PD | score\_for\_attribute |
| [1] MISSING | 37 |
| [2] LOW < - 0.098 | 117 |
| [3] 0.098 < - HIGH | 0 |

|  |  |
| --- | --- |
| CR\_AGE\_MARITAL | score\_for\_attribute |
| GROUP 1 | 0 |
| GROUP 2 | 16 |
| GROUP 3 | 23 |
| GROUP 4 | 67 |

|  |  |
| --- | --- |
| AGE\_IDCARD | score\_for\_attribute |
| [1] LOW < - 0.75 | 0 |
| [2] 0.75 < - 8.5 | 51 |
| [3] 8.5 < - HIGH | 93 |

|  |  |
| --- | --- |
| TIME\_WITH\_VPB | score\_for\_attribute |
| [1] NO RELATIONSHIP | 3 |
| [2] 0.07 < - 9 | 0 |
| [3] 9 < - 24.5 | 12 |
| [4] 24.5 < - HIGH | 40 |

|  |  |
| --- | --- |
| RISK\_ZONE2 | score\_for\_attribute |
| GROUP 1 | 60 |
| GROUP 2 | 0 |

Detailed score of each variable is presented in file “Definition\_upl\_pcb\_2018.xlsx”.

## Model statistic assessment

### Model discrimination

< Trình bày tính phân biệt của mô hình trên mẫu development, đối với mỗi tiêu chí có bao gồm bảng biểu, hình minh họa và đánh giá so với benchmark, giải thích nếu có>

Ví dụ: Model performance:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Score band | Total, # | Good, # | Bad, # | Bads% in band | %Bads in band & over |
| B1 | 871 | 675 | 196 | 22.5 | 100 |
| B2 | 846 | 732 | 114 | 13.5 | 81.1 |
| B3 | 940 | 829 | 111 | 11.8 | 70.2 |
| B4 | 797 | 714 | 83 | 10.4 | 59.5 |
| B5 | 900 | 826 | 74 | 8.2 | 51.5 |
| B6 | 858 | 791 | 67 | 7.8 | 44.4 |
| B7 | 2580 | 2434 | 146 | 5.7 | 37.9 |
| B8 | 837 | 802 | 35 | 4.2 | 23.9 |
| B9 | 2551 | 2446 | 105 | 4.1 | 20.5 |
| B10 | 932 | 900 | 32 | 3.4 | 10.4 |
| B11 | 741 | 720 | 21 | 2.8 | 7.3 |
| B12 | 883 | 860 | 23 | 2.6 | 5.3 |
| B13 | 753 | 743 | 10 | 1.3 | 3.1 |
| B14 | 1742 | 1725 | 17 | 1 | 2.1 |
| B15 | 806 | 801 | 5 | 0.6 | 0.5 |

Đánh giá:

### Model accuracy

< Trình bày tính chính xác của mô hình trên mẫu development, đối với mỗi tiêu chí có bao gồm bảng biểu, hình minh họa và đánh giá so với benchmark, giải thích nếu có>

# Pre-validation

## Model discrimination

<Tương tự trên, tuy nhiên đánh giá trên mẫu in-time validation hoặc OOT. Trong trường hợp không có mẫu 30% thì dùng bootstrapping để đáng giá. Cần mô tả quy trình thực hiện bootstrapping và kết quả.>

## Model accuracy

<Tương tự trên, tuy nhiên đánh giá trên mẫu in-time validation hoặc OOT. Trong trường hợp không có mẫu 30% thì dùng bootstrapping để đáng giá. Cần mô tả quy trình thực hiện bootstrapping và kết quả.>

## Model stability

<Mô tả tính ổn định của mô hình dựa trên tiêu chí đánh giá của toàn bộ mô hình và các biến cụ thể. Với mỗi tiêu chí bao gồm bảng biểu, hình minh họa và đánh giá tham chiếu với benchmark + giải thích nếu có>

**Ví dụ:** At the time of model development, the data provided by the PCB was only until Aug 2017, therefore, PSI was only calculated until this time. However, for variables from VPBank's internal information, data are collected until March 2018 to determine the stability of each variable. Analysis results show that:

* PSI is 12.2% for August 2017; it is moderately stable compared with the development model. However, this time is quite far from approval of the model; therefore, PSI after deployment can be higher.
* In the internal variables of VPBank, the Sale channel is the most unstable variable, however, according to Head of Retail Risk assessment, this is not a big problem. The other variables change between development sample and June 2017 sample from June 2017 and then keep stable to March 2018 without any major fluctuations.

## Model weakness/limitation/assumption

<Weakness & assumption>

## Plan of improvement

<Plans to improve model in the future>

# Model usage

## Model implementation

<Mô tả ít nhất các cấu phần sau

* Hệ thống dùng để implement mô hình
* quy trình vận hành/sử dụng mô hình trong thực tế.
* Đánh giá các sai khác nếu có so với dự kiến ban đầu, ví dụ các biến trong mô hình không được tính chính xác trên hệ thống.
* Đánh giá tính chính xác của dữ liệu dùng để chấm điểm trên thực tế
* Các rủi ro có thể gặp phải trong quá trình sử dụng mô hình>

## Responsibility of related party

<Mô tả vai trò, trách nhiệm của các bên liên quan>

## Related policy

<Mô tả các chính sách có dùng mô hình >

## Monitoring report

Model is monitored monthly to track PSI of whole model as well as component variables.

Detailed monitoring of the model over the months is shown in the two files “Monitoring\_upl\_pcb\_201708.xlsx” and “Monitoring\_VPB\_variables\_201803.xlsx” attached file in appendix.

# Model makers

<Mô tả trách nhiệm của từng cá nhân trong việc xây dựng mô hình theo file trách nhiệm đã có>

List of staff involved in model development

|  |  |  |
| --- | --- | --- |
| **Task** | **Main PIC** | **Support** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Table 18: List of participants

# Appendix

<Danh mục các tài liệu đính kèm:

* Biên bản họp/trao đổi với các bên liên quan trong quá trình xây dựng mô hình dưới dạng pdf
* File phân tích biến như đã yêu cầu
* File phê duyệt mô hình
* Biên bản UAT
* Các văn bản liên quan đến việc sử dụng mô hình
* Các báo cáo có sử dụng mô hình>
* Code sử dụng để lấy biến cho mô hình
* Code sử dụng để xây dựng mô hình

# Bibliography

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

[Bibliography includes: 1) external documents, 2) internal document, proposal… related to model development]