Five New Derived Variables

RatioPendingAmount

This variable gives the ratio of ArrearsAmount to DisbursedAmount. The variable tells what fraction of the money has not yet been paid from the total amount disbursed to the account holder. The variable will be expressed in percentage to make it easier to read.

RatioPendingAmount=(ArrearsAmount/`Disbursement Amount’n)\*100;

RatioInstallment

To determine whether a large monthly instalment has been deterrent in any way to repay the loan back. This value is given as a ratio of InstallmentAmt to total amount disbursed.

RatioInstallment=(InstallmentAmt/ `Disbursement Amount’n)\*100;

Arrears\_Delay\_Amount

If the upaid loan amount is large and it has been a long time since the account holder has not made any payment, then it is likely that it would be a bad loan. Hence those loans where unpaid balance is low and arrears day is large, or loans where the unpaid balance is high, but arrears day is low is less likely to be a risky loan.

To normalize the unpaid loan amount, ActualBalance will be divided by the amount disbursed and ArrearsDays will be normalized as years. After normalization both the variables are the multiplied.

Arrears\_Delay\_Amount = (ActualBalance/`Disbursement Amount’n) \* (ArrearsDays/365);

RatioArrearsDay

This is the ratio of ArrearsDay to ‘Days on File’ to determine for what percentage of time the account holder has been defaulting the payment while keeping in consideration the time period for which the account holder timely paid the loan instalments.

RatioArrearsDay = (ArrearsDay/‘Days on File’n)\*100;

ArrearsDay\_Rank

This variable determines the risk level of “bad loan”. It is calculated as follow.

If ArrearsDays<35 then ArrearsDayRank=`Low Risk’;

Else if ArrearsDays >=35<338 then ArrearsDayRank=`Medium Risk’;

Else if ArrearsDays >=338 then ArrearsDayRank=`High Risk’;

Where for the variable ArrearsDay Q1=35, and Q3=338.

Commentary of variable

AccountID

This variable uniquely identifies the account. A “bad loan” will be independent of the Account ID, hence it will be dropped from my final clean data file.

BranchID

This variable indicates the branch which issued the loan to an applicant. The variable’s accuracy, consistency and completeness are all high. The variable will be recoded to make it easier to read. In my analysis I will evaluate whether there is any correlation between “bad loan” and the branch which issued the loan.

2='Mandala'

3='Kawale'

4='Mzuzu'

5='Blantrye'

6='Salima'

7='Mchinji';

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 0 | 0 | BranchID 7 | 42 | BranchID 2 | 680 |

ProductID

This variable identifies the type of product/ loan given to the account holder. There are no missing values for this variable. This variable will be used as an input variable to determine whether people tend to default on certain categories of loan product.

The variable’s accuracy, consistency and completeness are all high. The variable will be recoded to make it easier to read.

'LN01'='Group Business Loan'

'LN02'='Individual Business Loan'

'LN03'='Loan to purchase Shares'

'LN04'='Emergency Loan'

'LN05'='Farming Loan'

'LN06'='Woman Loan'

'LN08'='Payroll Secured Loan';

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 0 | 0 | LN05 | 48 | LN08 | 447 |

DisbursedOn

This is a date variable when the loan was disbursed to accountholder. This variable is valuable to define new variable which can more information regarding for what fraction of time payment has been under arrears.

The variable’s accuracy, consistency and completeness are all high. No changes in the values of the variable will be performed.

ApplicationScore

A higher score indicates that the branch ‘predicted’ that this application would be minimal risk of defaulting on the loan. A low score indicates that the branch ‘predicted’ that the loan was at a higher risk of defaulting. This is a target variable and will be used to evaluate the independent variables.

The variable’s accuracy, consistency and completeness are all high. There are no outliers, hence no change in the value of the variable will be performed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Min** | **5th Percentile** | **50th Percentile** | **95th Percentile** | **Max** |
| **0** | **0** | **9** | **13** | **26** | **42** | **52** |

ArrearsDays

The number of days that the loan is in arrears (i.e. behind in its payments). The current date of this file extract was July 31, 2012. This is a target variable and will be used to evaluate the independent variables.

An extreme value is given as a values which are higher than Q3+3\*IQR, where IQR is the difference between Q3 and Q1. There are four such instances which will be removed.

The variable’s accuracy, consistency and completeness are all high.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Min** | **5th Percentile** | **50th Percentile** | **95th Percentile** | **Max** |
| **0** | **0** | **1** | **5** | **134** | **685** | **1847** |

Actual Application Grade

Letter grade of loan application quality. The values will be renamed, so that meaningful names are used instead of using letter grades. The variable’s accuracy, consistency and completeness are all high. The following convention will be used:

'A'='Best Quality'

'B'='Good Quality'

'C'='Average Quality'

'D'='Poor Quality'

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 0 | 0 | Grade A | 121 | Grade C | 676 |

Actual Good Bad

Internal indicator where 1 – account is considered ‘Good’, and 0 – account is considered ‘Bad’. The values can be recoded to make it easier to read the data values. To make the variable name more meaningful it would be renamed as Internal\_Indicator.

The variable’s accuracy, consistency and completeness are all high.

1='Good account'

0='Bad account'

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 0 | 0 | 1 | 805 | 0 | 807 |

Gender

The gender of the individual who was given the loan, when the loan is given to a group of individuals it is marked as "Group".

There are six missing values which would be assigned the value of “group” as part of the data cleaning process. The variable’s accuracy, consistency and completeness are all high. Recode M and F to make it easier to read the values.

'M'='Male'

'F'='Female'

'G'='Group'

' '='Group'

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 6 | 0.37 | Group | 214 | Male | 965 |

LiteracyLevel

The education level of account holder. D and H are outside the range. And U refers to Unspecified and University education, hence it cannot be inferred what does U refers to. Similar values should be grouped together.

'X'='Not Specified'

'D'='Not Specified'

'H'='Not Specified'

'N'='No Formal Education'

'O'='No Formal Education'

'P'='Primary'

'E'='Primary'

'C'='Tertiary'

'PT'='Tertiary'

'T'='Tertiary'

'U'='Tertiary'

'S'='Secondary'

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 0 | 0 | H | 6 | E | 867 |

Occupation

The profession of the individual. Several of the values can be merged together, for example, B which represents "Business People" and M which represents "Self Employed" can both be grouped as "Business person".

'PR'='Employee'

'E' ='Employee'

'PRO'='Civil Servants'

'B'='Business-person'

'T'='Business-person'

'M'='Business-person'

'C'='Civil Servants'

'F'='Farmer'

'N'='Unemployed'

' '='Other'

'O'='Other'

'U'='Other';

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 1 | 0.06 | N | 2 | B | 736 |

MaritalStatus

The marital status of the account holder. The value "P" is out of range. The values "widow" and "widower" can be merged. Blank values can be assigned a value of "Not specified". This variable will be used as an input variable to determine whether account holders of certain marital status are more likely to default a loan.

'D'='Divorced'

'G'='Group'

'M'='Married'

'S'='Single'

'U'='Not Specified'

'P'='Not Specified'

' '='Not Specified'

'W'='Widow/er'

'WI'='Widow/er';

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 5 | 0.31 | Divorced | 7 | Married | 1037 |

PurposeCode

The purpose for which the loan was obtained. Currently there are many possible categories in this field. These values will be clustered in small number of clusters with similar values grouped together. Blank values will be assigned a value of "Other".

This variable will be used as an input variable to determine whether account holders of borrowing loan for a certain purpse are more likely to default a loan.

1='For Business'

5='For Business'

11='For Business'

4='Bills Payment'

6='Bills Payment'

8='Bills Payment'

9='Bills Payment'

12='Bills Payment'

2='Real Estate'

3='Real Estate'

7='Real Estate'

10='Real Estate'

.='Other'

13='Other';

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 3 | 0.19 | 4 | 12 | 1 | 624 |

DonorID

Many values are missing, and there is little variance in the data. More than 88% of loan money is coming from a single donor. Also, a "bad loan" is unlikely to be affected by who has provided the donation, hence this variable will be dropped from the clean data file.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Minimum Frequency Category** | **Minimum Frequency** | **Maximum Frequency Category** | **Maximum Frequency** |
| 125 | 7.75 | DonorID 202 | 1 | DonorId 200 | 1420 |

LoanSeries

No information is available to interpret what is meant by LoanSeries, hence this variable will be dropped during the cleaning process.

CreditOfficerID

A “bad loan” is independent of which credit officer handled an applicant’s application, hence this variable will be dropped during the cleaning process.

SectorID

No information is available to interpret what is meant by LoanSeries, hence this variable will be dropped during the cleaning process.

Disbursement Amount

The amount of loan which was disbursed to the applicant. As the loan is given by a small micro finance bank for small amount of loan, very large loan values (>310000, 95 percentile) would be treated as outlier and removed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Min** | **5th Percentile** | **50th Percentile** | **95th Percentile** | **Max** |
| 0 | 0 | 1000 | 5000 | 40000 | 310000 | 1605000 |

InstallmentAmt

The required monthly payments on the loan. As the loan is given by a small micro finance bank for small amount of loan, very large instalment values (>49170, 95 percentile) should be treated as outlier and removed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Min** | **5th Percentile** | **50th Percentile** | **95th Percentile** | **Max** |
| 0 | 0 | 150 | 920 | 5500 | 49170 | 322050 |

ActualBalance

The current balance of all unpaid portions of the loan. Remove extreme outliers, given by > 172104, 95 percntile.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Min** | **5th Percentile** | **50th Percentile** | **95th Percentile** | **Max** |
| 0 | 0 | 0 | 2587 | 21249 | 172104 | 1500000 |

ArrearsAmount

The current amount of all loan payments that have not yet been paid (are past due).

Remove extreme outliers (>464179, 99 percentile).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **# Missing** | **% Missing** | **Min** | **5th Percentile** | **50th Percentile** | **95th Percentile** | **Max** |
| 0 | 0 | -128000 | 0 | 16237 | 190167 | 1753770 |