5.1 Performance testing

5.1.1 Data quality

In data quality analysis, we focussed on assessing whether the data used in input of the model (e.g., data from systems and databases) are reliable. Following tests are performed: 1) Missing values 2) Descriptive statistics

Missing values

|  |  |
| --- | --- |
| Feature | Missing |
| mrfytdocc | 42.87% |
| priorfyocc | 25.28% |
| priorfydscr | 21.82% |
| debt\_yield\_p1 | 17.51% |
| OLTV | 3.4% |

Descriptive statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metrics | Priorfydscr | Mrfytdocc | Debt yield p1 | Priorfyocc | Oltv | Division | Interestonly | Bad flag final v3 |
| count | 72687 | 53116 | 76691 | 69471 | 89812 | 92974 | 92974 | 92974 |
| unique | N/A | N/A | N/A | N/A | N/A | 10 | 3 | N/A |
| top | N/A | N/A | N/A | N/A | N/A | South-Atlantic | N | N/A |
| freq | N/A | N/A | N/A | N/A | N/A | 19311 | 85294 | N/A |
| mean | 1.8395755298 | 93.4809228426 | 12.5337343705 | 93.754880108 | 66.4377253596 | N/A | N/A | 0.011960333 |
| std | 1.8170684715 | 9.0892893303 | 13.8366159547 | 9.2084476055 | 17.2802191395 | N/A | N/A | 0.1087079139 |
| min | -2.8302 | 0.78 | -26.38654378 | 1 | 0.9 | N/A | N/A | 0 |
| 25% | 1.2978 | 91.9175 | 6.8698379192 | 92 | 62.8 | N/A | N/A | 0 |
| 50% | 1.53 | 95.1 | 10.461944328 | 96 | 72.59 | N/A | N/A | 0 |
| 75% | 1.86 | 100 | 14.0404936145 | 100 | 77.6 | N/A | N/A | 0 |
| max | 63.97 | 100 | 408.3590591 | 100 | 96.24 | N/A | N/A | 1 |

5.1.2 Information value

Information Value (IV), calculated at risk driver level, is a measure used to quantify the predictive power of a risk driver. More specifically, how well can a risk driver discriminate between a default and no-default response.

The test returns IV for selected risk drivers listed in the table below.

|  |  |
| --- | --- |
| Feature | Iv total |
| priorfydscr | 0.8183686365 |
| priorfyocc | 0.5424667177 |
| mrfytdocc | 0.409471795 |
| debt\_yield\_p1 | 0.293684185 |
| OLTV | 0.2461692548 |
| Division | 0.235608834 |
| interestonly | 0.0443585828 |

The table above provides the IV of all the variables used in the model, higher IV implies more discriminatory power of the variable between the default and non-default customers.

5.1.3 Model coefficient

The PD model is developed using logistic regression technique with 7 risk drivers (independent variables) in the final model equation. Please find below the list of parameters and their model coefficients.

|  |  |
| --- | --- |
| Parameter | Coef |
| const | -4.4168461324 |
| priorfydscr | -0.7618294827 |
| mrfytdocc | -0.3871917877 |
| debt\_yield\_p1 | -0.1510340457 |
| priorfyocc | -0.4915827429 |
| OLTV | -0.310497772 |
| Division | -0.5789169907 |
| interestonly | -0.3459055608 |

5.1.3.2 Scoring

Based on the forecasted probability of default, each facility is assiged to a rating. Below the distribution of the ratings for the train and test dataset.

Train dataset distribution

Test dataset distribution