5.1 PD model replication

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 Univariate analysis

In a univariate analysis, we set up a multitude of logistic regression models where there is only one explanatory variable (the risk driver) and the response variable (the default). Among those models, the ones that describes best the response variable can indicate the most significant explanatory variables, which can then be used to perform a multivariate analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Coef | P z | Gini coefficient |
| priorfydscr | -0.881726911 | 0 | 0.452 |
| mrfytdocc | -1.2194863213 | 0 | 0.321 |
| debt\_yield\_p1 | -0.8919738737 | 0 | 0.281 |
| priorfyocc | -1.1645963421 | 0 | 0.383 |
| OLTV | -0.7754851517 | 0 | 0.217 |
| Division | -0.8479998614 | 0 | 0.238 |
| interestonly | -0.9347473414 | 8.31589183e-158 | 0.057 |

The model coefficients, p-values and Gini coefficients of each of the chosen explanatory variables indicate that these parameters explain the reponse variable quite well.

5.1.3 Model coefficient comparison

The PD model is independently replicated by the validation team to ensure correctness of the implementation. In order to assess whether the replicated PD model yields the same output as the original model, the risk drivers (independent parameters) and their model coefficients are compared.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Coef development | Coef validation | Coef diff |
| const | -4.4168461324 | -4.4168461324 | 0 |
| priorfydscr | -0.7618294827 | -0.7618294827 | -2.220446049e-16 |
| mrfytdocc | -0.3871917877 | -0.3871917877 | -9.992007222e-16 |
| debt\_yield\_p1 | -0.1510340457 | -0.1510340457 | 0 |
| priorfyocc | -0.4915827429 | -0.4915827429 | -5.551115123e-17 |
| OLTV | -0.310497772 | -0.310497772 | -8.881784197e-16 |
| Division | -0.5789169907 | -0.5789169907 | 0 |
| interestonly | -0.3459055608 | -0.3459055608 | 0 |

The differences between the results of the independent model replication and the results of the development team, both for the model coefficients and their standard deviation, indicate the correctness of the implementation of the PD model. In case these differences are not equal to zero, further investigation is needed into why this would be the case (e.g., implementation issues, specifications of the underlying data set, parameterization of the PD model, etc).

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