Rethinking Representativeness Analysis in IRB Modeling

Classifier Two-Sample Test

Andrija Djurovic

www.linkedin.com/in/andrija-djurovic

Representativeness in IRB Modeling

- Different regulatory documents emphasize that the data used for model development and calibration under the IRB framework must be representative of the institution's actual portfolio.
- Moreover, practitioners are expected to analyze representativeness across different dimensions. These dimensions are often divided into qualitative and quantitative categories, with the distribution of risk factors being one of the main quantitative aspects of representativeness analysis.
- In practice, this involves statistical comparisons of risk factor distributions across different modeling datasets and time periods.
- Given these requirements, practitioners often face challenges in forming an overall conclusion on the representativeness of risk factor distributions.
- Currently, the industry standard for testing changes in the distribution of risk factors is the Population Stability Index (PSI), with commonly used thresholds of 0.10 and 0.25 indicating low, moderate, and high shifts. However, since PSI is applied at the level of individual risk factors and requires discretized inputs for its calculation, practitioners often face challenges in drawing an overall conclusion from this analysis.
- In this presentation, Andrija Djurovic proposes an alternative approach based on the Classifier Two-Sample Test (C2ST). Although primarily considered in the field of Machine Learning (ML), to the best of the author's knowledge, C2ST has not yet been explored in the context of IRB modeling. As C2ST can address some of the main weaknesses of PSI testing, the author recommends considering it as part of the standard representativeness analysis, alongside other methods, to support more comprehensive conclusions.

Classifier Two-Sample Test

The main goal of the Classifier Two-Sample Test (C2ST) is to assess whether two samples are representative of each other. In the context of IRB modeling, one example is to test whether the train dataset is representative of the application portfolio to which the model is intended to be applied.

Let us assume that practitioners have a train and a test dataset available, where the train dataset represents the data used to develop the risk differentiation function, and the test dataset represents another sample for which representativeness needs to be evaluated. In IRB modeling, this can refer to an out-of-sample, out-of-time, or application dataset, or any other dataset used for this purpose. Under the assumption that these two samples are prepared, the following steps illustrate the process of testing representativeness between them:

1. Merge the training and test datasets and create an indicator variable equal to 1 for one sample and 0 for the other.
2. Select the risk factors for which changes in distribution are to be tested.
3. Choose any classification or regression algorithm suitable for a binary problem.
4. Run the selected algorithm.
5. Calculate the Area Under the Receiver Operating Characteristic Curve (AUC ROC) and test whether it is statistically different from 0.5. If the AUC ROC is statistically different from 0.5, conclude that there is evidence that the train sample is not representative of the test sample.

Classifier Two-Sample Test cont.

Some of the these steps may vary depending on the purpose and stage of the representativeness analysis. Furthermore, suppose C2ST reveals that the train dataset is not representative of the test dataset. In that case, practitioners can apply feature importance techniques to identify which risk factors potentially drive the difference in AUC ROC.

As described in the procedure, this test is based on statistical inference rather than arbitrary thresholds. It can handle mixed types of risk factors (continuous and categorical), simultaneously assess distributional changes across all variables, and identify individual risk factors contributing to the issue.

However, despite these advantages, since the method relies on statistical testing, its conclusions may be influenced by sample size. Therefore, it is recommended to combine this approach with other methods. The author particularly suggests combining it with the model shift concept developed by Dr. Alan Forrest. Basic details about the model shift concept can be found in the following link.

Simulation Study

Assume that the final PD model is developed on the train dataset, where the target variable is given in the column `Creditability`, while the remaining five columns represent the risk factors selected in the final model. Furthermore, assume that the test dataset is an out-of-time sample used to assess the representativeness of the selected risk factors using the classifier two-sample test (C2ST) approach.

To conduct the representativeness testing using the C2ST approach, the following steps are first executed:

1. Encode all risk factors in the train dataset using the Weight of Evidence (WoE) method.
2. Using the WoE values calculated on the train dataset, encode the risk factors in the test dataset.

Once the above steps are completed, the C2ST can be run as follows:

1. Merge the train and test datasets and add a new variable named indicator, which takes the value 0 for all observations from the train dataset and 1 for all observations from the test dataset.
2. Run a logistic regression where the indicator variable is the dependent variable, and all other variables serve as predictors.
3. Calculate the AUC ROC for the model from step 2.
4. Test whether the AUC ROC is equal to 0.50.
5. If the AUC ROC is statistically different from 0.50, run a permutation-based importance procedure to identify the risk factors that drive the conclusion that the train dataset is not representative of the test dataset.

Simulation Study cont.

Practitioners should note that the described procedure serves only as a demonstration and can be substantially adjusted depending on the stage of testing and the choice of binary classification algorithm. Therefore, practitioners are encouraged to adapt these steps to their specific needs.

Furthermore, even after identifying risk factors that may indicate non-representativeness, the author recommends a thorough analysis of their potential impact on the risk differentiation function and final estimates.

The following slides present the results of the simulation study described. For the sake of reproducibility, specific intermediate results are also shown. The AUC ROC testing approach proposed by Hanley and McNeil (1982) is implemented.

Simulation Results

The following table provides an overview of the PD model with WoE encoding for categorical risk factors.

```
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -0.8496     0.0904 -9.3951   0.0000
## `Account Balance`                    -0.9022     0.1079 -8.3610   0.0000
## `Payment Status of Previous Credit`  -0.7800     0.1620 -4.8160   0.0000
## `Duration of Credit (month)`         -0.5601     0.1975 -2.8356   0.0046
## `Credit Amount`                      -0.7717     0.2084 -3.7030   0.0002
## `Age (years)`                        -0.7528     0.2407 -3.1272   0.0018
```

After preparing the train and test datasets for the C2ST analysis, the following table presents an overview of the logistic regression, using the same WoE encoding as in the PD model, but applied to a 0/1 indicator defined based on the datasets used.

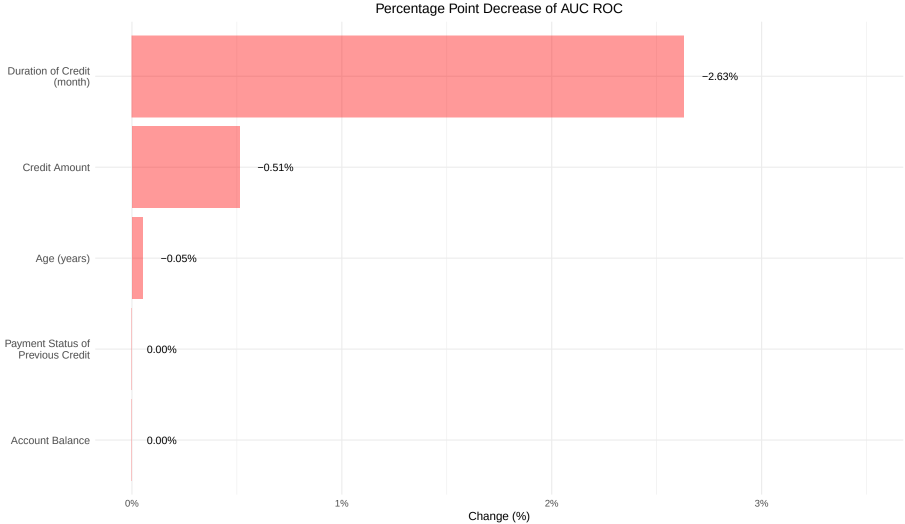
```
##                                     Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                          -1.4112     0.0826 -17.0906   0.0000
## `Account Balance`                     0.0822     0.0863   0.9524   0.3409
## `Duration of Credit (month)`         -0.4198     0.1799  -2.3338   0.0196
## `Payment Status of Previous Credit`  -0.0976     0.1452  -0.6724   0.5013
## `Credit Amount`                       0.2519     0.2053   1.2270   0.2198
## `Age (years)`                         0.2055     0.2333   0.8807   0.3785
```

Given the C2ST regression, the AUC ROC is calculated and tested against a value of 0.50.

```
##   AUC Observed AUC Test AUC Standard Error Test Stat. p-value
## 1       0.5609      0.5             0.0232     2.6293  0.0086
```

Since the calculated p-value is less than the typically chosen significance level of 5%, it suggests that there is evidence that the train dataset is not representative of the test dataset. To further investigate which risk factors may impact this conclusion, the following slide presents the results of the permutation-based approach for feature importance.

Simulation Results cont.


Percentage Point Decrease of AUC ROC
Duration of Credit (month)
Credit Amount
Age (years)
Payment Status of Previous Credit
Account Balance
Change (%)