Machine Learning and Credit Risk: Empirical Evidence from SMEs

Alessandro Bitetto [1] Paola Cerchiello [1] Alessandra Tanda [1] Barbara Tarantino [1] Stefano Filomeni [2]

[1] University of Pavia, Italy

[2] University of Essex, UK

Big Data and Machine Learning in Finance Conference
Milan (online)
June 11th, 2021

Overview

Conceptual framework

Credit scoring has experienced a natural evolutionary advancement using *machine-driven classification tools* that enables insurers to develop more accurate algorithms to assess creditworthiness. At the same time, these algorithms have been criticized for being *black boxes*.

Problem definition and objectives

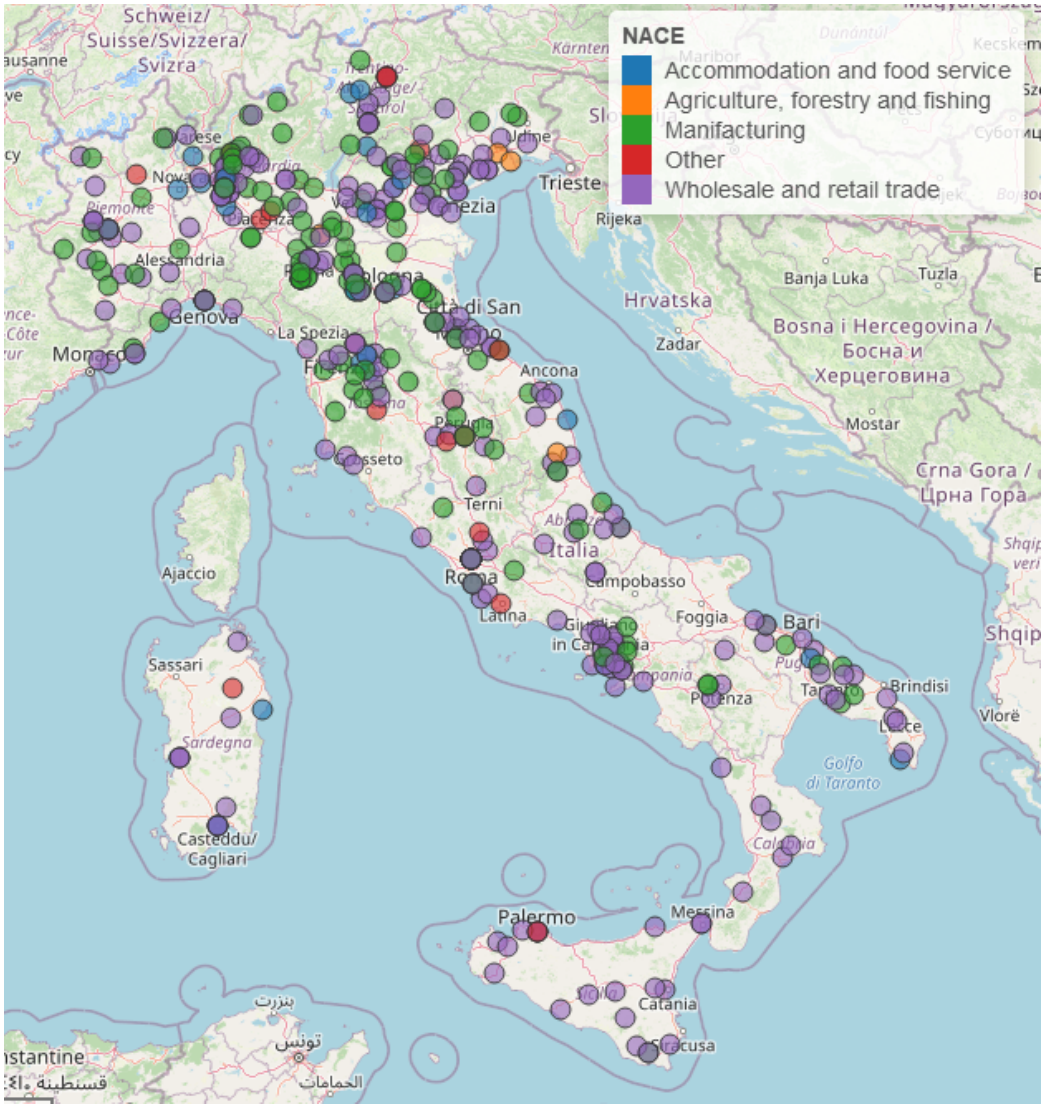
Methodological extension has been presented to examine static and dynamic modeling framework for credit risk assessment in insurance companies.

- Investigate the **determinants of credit rating** as a function of financial and business variables
- Examine **persistence in rating process** with regards to internal rating models in comparison to static modeling framework
- Implement general model-agnostic methods for **interpreting *black box* models** as Permutation Feature Importance and Shapley values

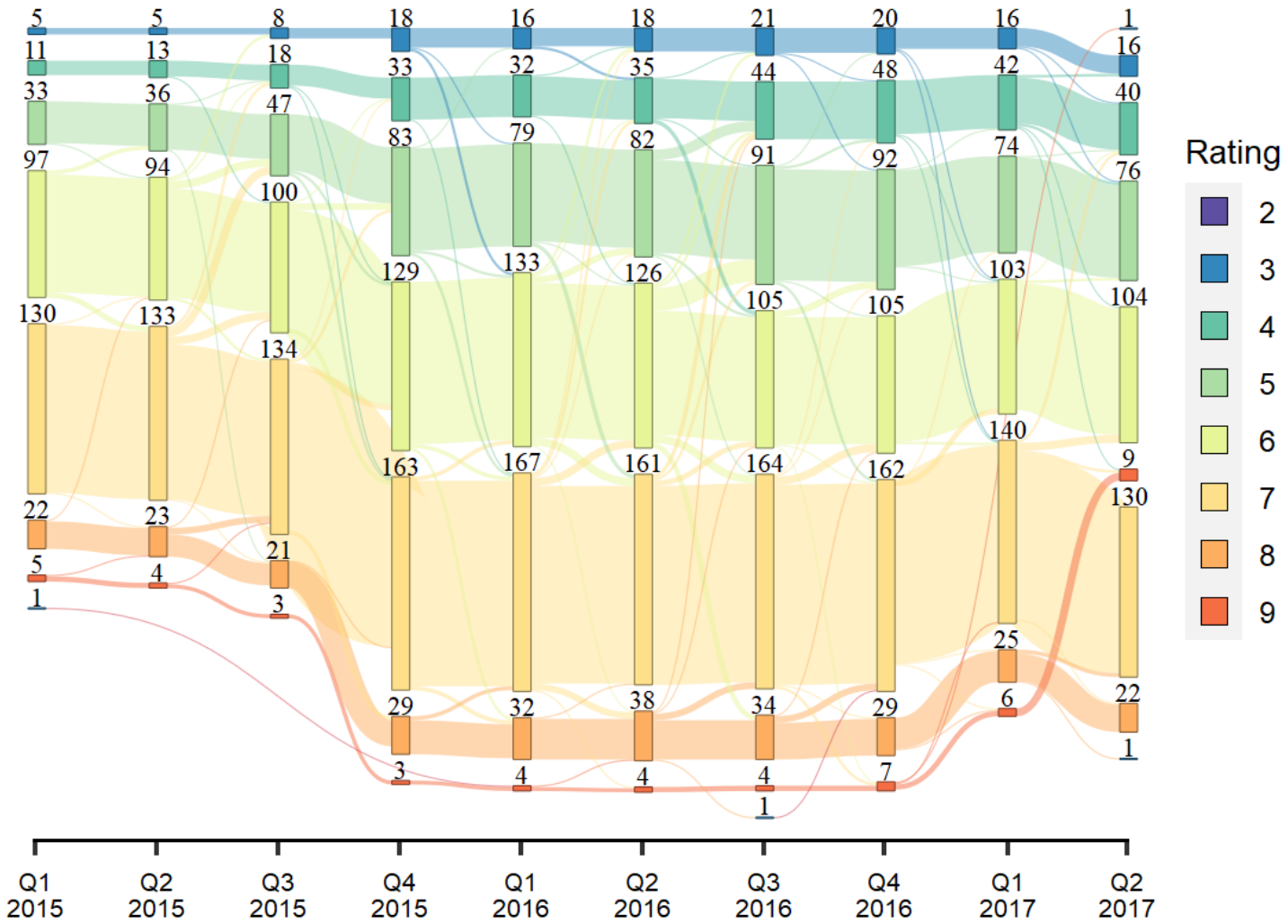
Dataset

- A panel of 810 Italian Firms in the time period that spans from the first quarter of 2015 to the second quarter of 2017 with quarterly frequency
- The target variable of interest is the *long-term rating* assigned to each firm, categorized in a numeric scale from *2* to *9* according to their credit risk. The higher the number, the worse the score.
- Among the predictors :
  - 23 balance sheet fundamentals variables (total asset, liability, ROE, etc) with annual frequency, taken from Orbis
  - 6 securitization variables (collection, oustanding, delinquency, etc) with quarterly frequency, given by the insurance company
  - 3 geographical and sector variables (region, economic sector, industry), taken from Orbis

Data heterogeneity



Rating distribution over time



Methodology

- Data pre-processing results in a final dataset of 20 variables and 534 firms
- Analysis is splitted in a combination of three dimensions :
  - Set of variables : *Fundamentals* (Orbis) vs *Securitization* (Insurance)
  - Temporal dynamics : *Static* vs *Dynamic*
  - Models : *Ordered Probit* vs *Historical Random Forest*
- All models are calibrated using hyperparameter tuning (with Bayesian Optimization) and Rolling Window Temporal Cross-Validation
- Permutation Feature Importance and Shapley values are evaluated to assess feature importance

Ordered Probit Model

Static modeling framework

The ordered probit model is defined as :

$$y_{it} = X_{it}\beta + \alpha_i + \epsilon_{it}$$

where y_{it} is the target variable for firm i at time t, X_{it} is the set of variables, β is the set of parameters to be estimated, α_i is a firm-specific and time-invariant component and ϵ_{it} is error term

Dynamic modeling framework

In the dynamic version, two more terms are added :

$$y_{it} = X_{it}\beta + y_{it-1}\gamma + y_{i0}\delta + a_i + \epsilon_{it}$$

where y_{it-1} indicates a vector of firm's rating in the previous quarter, γ denotes the parameters linked to rating in the previous quarters, y_{i0} is the value of the dependent variables in the initial period.

Historical Random Forest

Static modeling framework

The static version of the model is the standard Random Forest model. Each target variable y_{it} is associated with the concurrent input variables X_{it}.

Dynamic modeling framework

The dynamic version of the model adds summary variables of each X_i as an aggregation of all observed values over time. Possible summary functions $S(\eta; z_{ijk})$ for observation z_{ijk} are :

- frequency normalization

$$\sum_{t_{ij}-\eta_1 \leq t_{ih} < t_{ij}} \frac{I(z_{ihk} < \eta_2)}{n_{ij}(\eta)}$$

- average

$$\sum_{t_{ij}-\eta_1 \leq t_{ih} < t_{ij}} \frac{z_{ihk}}{n_{ij}(\eta)}$$

where I is the counting function and n_{ij} is the number of elements for observation i in time window $[t_{ij} - \eta_1, t_{ij} - \eta_2]$

Model tuning

- The type of resampling is k-fold cross-validation using *rolling origin fixed window* subsamples (with k = 4). The origin is updated each time, adding one new observation to the fit period. A fixed window (i.e. fit period) of constant length has been maintained while building folds.
- Optimal model architecture has been selected based on a weighting adjustment on the F_1-score aimed at maximizing F1 score on validation set and minimizing the distance between F1 score on training and validation set.

$$F_{1\,\mathrm{w}} = -F_{1\,\mathrm{test}}^{\gamma}\log(1 - F_{1\,\mathrm{test}}) - (1 - \Delta F_{1\,\mathrm{train\text{-}test}})^{\gamma}\log(\Delta F_{1\,\mathrm{train\text{-}test}})$$

where $\gamma \geq 1$.

Model explainability

Shapley Values

- Shapley values represent the marginal contribution of each feature to the prediction of a given data point, given by the difference between the feature effect and the mean effect.
- A Monte-Carlo sampling was proposed by Strumbelj et al.(2014) :

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^{M} (\hat{f}(x_{+j}^m - \hat{f}(x_{-j}^m)) \tag{1}$$

where $\hat{f}(x_{+j}^m)$ represents the prediction for the istance of interest x but with a random permutation of features (taken from a random data point z) except for jth feature. The vector x_{-j}^m is identical to x_{+j}^m, but also the value for feature j is randomized from the sampled z.

Model evaluation

Model	Version	Sample	F1 score
Probit	Static	Train	0.4579
		Test	0.4284
	Dynamic	Train	**0.7901**
		Test	**0.7346**
HRF	Static	Train	0.4105
		Test	0.3417
	Dynamic	Train	0.748
		Test	0.5519

TABLE – Securitization set

Model	Version	Sample	F1 score
Probit	Static	Train	0.4634
		Test	0.4529
	Dynamic	Train	0.8148
		Test	0.7407
HRF	Static	Train	0.919
		Test	0.677
	Dynamic	Train	**0.9154**
		Test	**0.7361**

TABLE – Fundamentals set

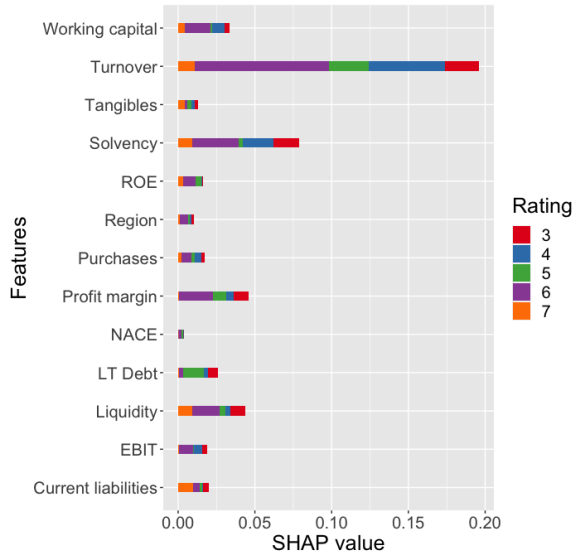
Model	Version	Sample	F1 score
Probit	Static	Train	0.4609
		Test	0.4543
	Dynamic	Train	0.799
		Test	0.74487
HRF	Static	Train	0.9611
		Test	0.6986
	Dynamic	Train	**0.9014**
		Test	**0.7326**

TABLE – Both sets

Shapley values

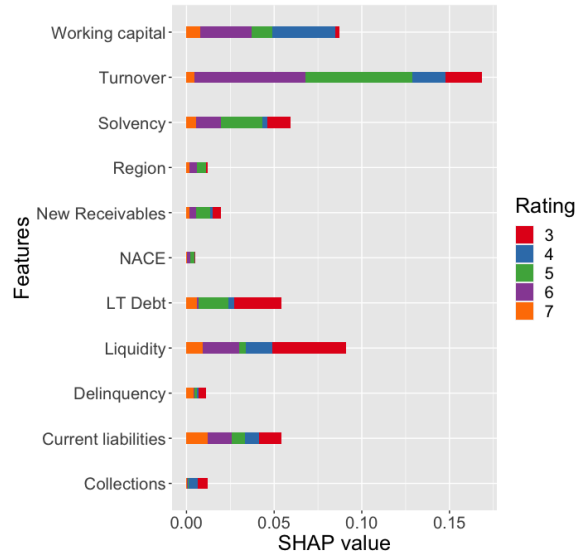
(a) Shapley values for securitization set - HRF
Top features :

- Delinquency 90
- Delinquency



(b) Shapley values for fundamentals set - HRF
Top features :

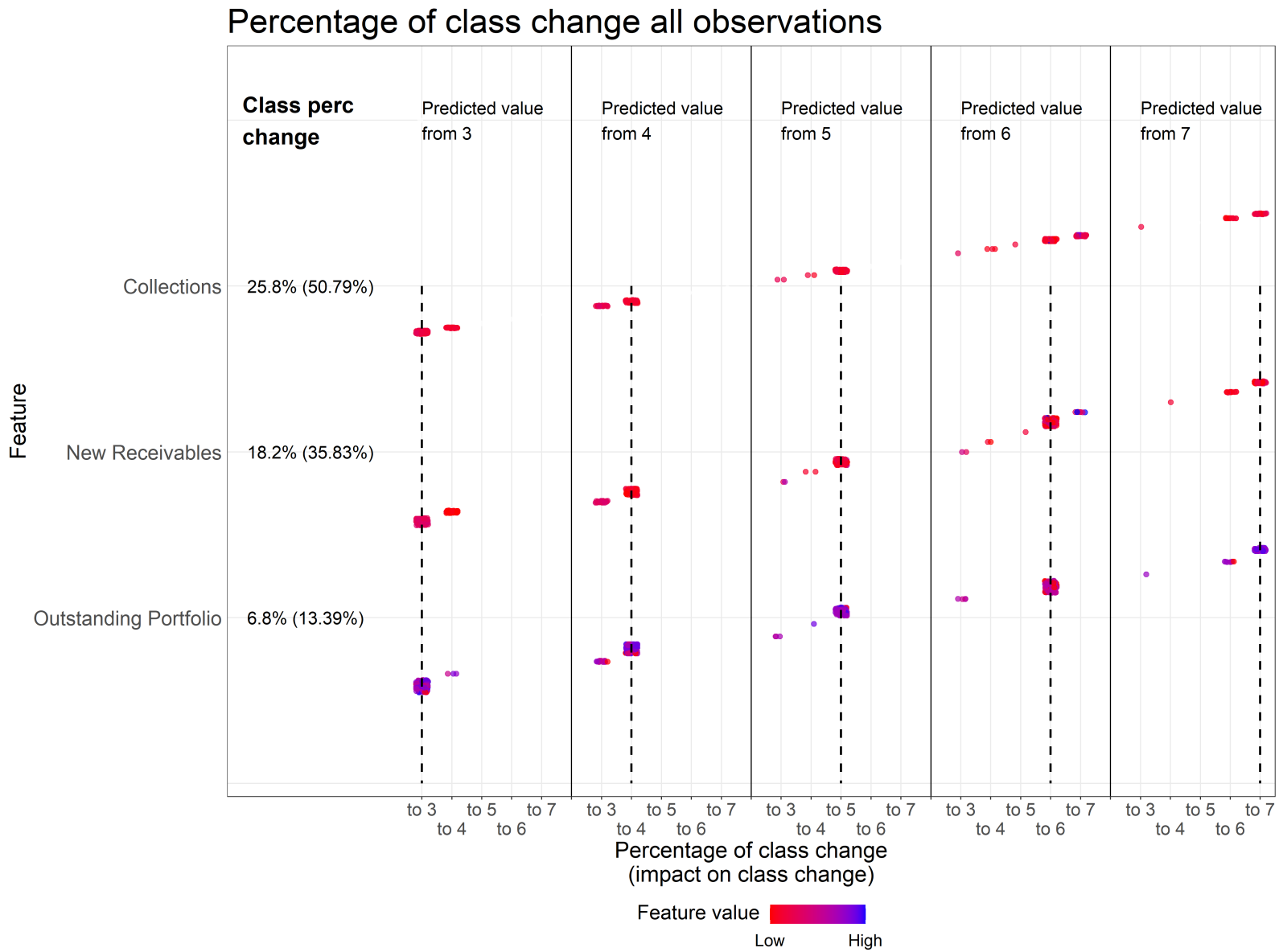
- Turnover
- Solvency



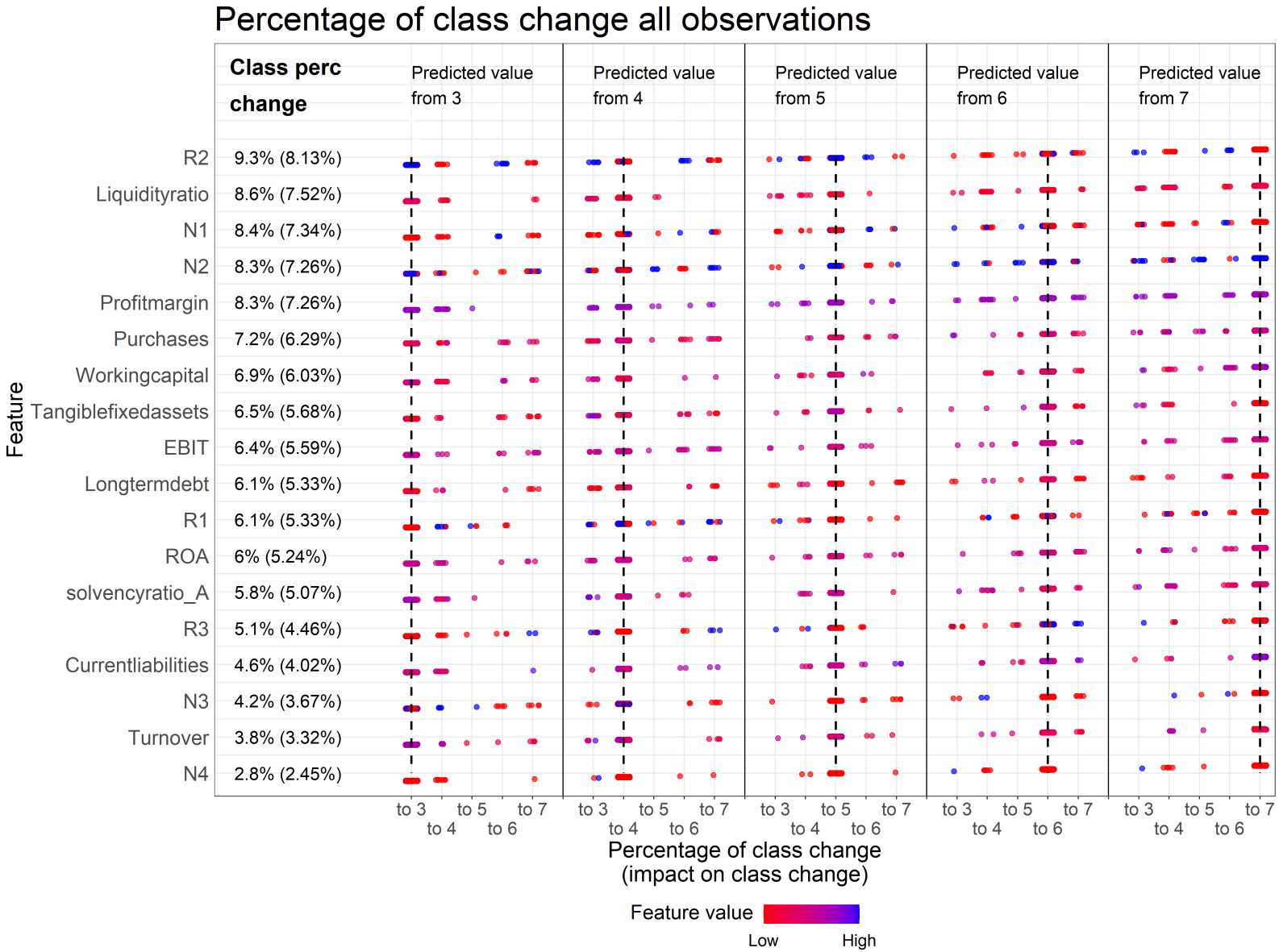
(c) Shapley values for both sets - HRF
Top features :

- Turnover (F)
- Liquidity (F)
- Working capital (F)

Shapley values : changes of class for Securitization set



Shapley values : changes of class for Fundamentals set



Assessment of differences and robustness

Statistical comparison of classifiers

- The p-values matrix generated from Wilcoxon signed-ranks test and post hoc tests for Friedman with regards to pairwise combinations highlighted the temporal component as statistically significant discriminant between algorithms.

Alternative formulation of the target variable

- The ordinal scale has been converted to dichotomous variable, where class 6 and 7 represent High-risk Grade against the other classes.
- The binary formulation results in slightly higher performances together with same explainability conclusions as for individual risk.
- Duality of sign is reported from partial derivatives since the threshold that highlights the change of sign is class 6.

Conclusions and future works

- The dynamic component of models is necessary to capture the temporal persistence of ratings and reduce misclassification cost.
- The good performance of Historical Random Forest algorithm (90% for train and 75% for test) is in line with other results in literature and seems to be a good choice for machine learning applications in credit risk.
- A similar approach used in our previous work may be tested to create a data-driven ranking of the firms riskiness and compared with the insurance rating