Why not just using a single indicator?

Falco J. Bargagli Stoffi

29/02/2020

Introduction

In this section of the Appendix we check what is the comparative advantage of using our Machine-Learning-based definition of Persistently Distressed Firms (PDF) rather than other commonly used deterministic indicators (i.e., Interest Covarage Ratio [ICR], negative added value, etc.).

The first and more obvious advantage is that the PDF indicator can be constructed even in the presence of missing data. As we show in the Section of the paper about missing data, there are significant missingness patterns in the two variables that compose the ICR indicator (EBIT and Interest Paid).

The second advantage is that, while ICR and the other indicators are based on a deterministic definition of zombies, PDF is based on a probabilistic definition that can be tuned to embody different "likelihoods" of being zombie.

The third advantage is that our PDF indicator can be constructed using the information coming from multiple indicators and not just using the information on a single one.

Beside these clear advantages, there are more subtle benefits that can be obtained by using PDF instead of ICR. To highlight these advantages we focus on a particular dimension of our PDF definition. To be a PDF a firm needs to be at high risk of failure (but not failing) for three consecutive years. Our PDF definition focuses on those firms that "are predicted to fail, but are surviving" (following the ML literature we could call them *false positives*). Hence, a good indicator of "*zombieness*" should be able to discriminate between *false positives* and *true positives*.

In this spirit, we develop a two stage algorithm to highlight which are the best predictors of *false positives*. In the first stage, we train a Logit model to predict the probability of failure:

$$f_{LOGIT}(x) = \hat{p}_i(Y_i = 1|X_i = x), \tag{1}$$

and to obtain the fitted values $\hat{y}_i$. In the second stage, we fit a LASSO just on those observations with $\hat{y}_i = 1$ (*positives*) to get the most important predictors (namely, those with non-zero coefficient) for the *false positives*.

In the following code chuncks the implementation of this "two-stage" algorithm.

Load Packages and Data

In this chunk we load the packages and the data used for the analysis, and we split the overall sample in a *training* and a *test* set.

```
options(java.parameters = "-Xmx50g")
library(rJava)
library(bartMachine)
library(haven)
library(plyr)
library(dplyr)
library(PRROC)
library(caret)
library(glmnet)
```

```r
# Load Data
data <- read_dta("analysis_data_indicators.dta")

# Initialize Data
names(data)[names(data) == 'GUO___BvD_ID_number'] <- 'guo'
data$control <- ifelse(data$guo=="", 0, 1)
data$nace <- as.factor(data$nace)
data$area <- as.factor(data$area)
levels(data$nace) <- floor(as.numeric(levels(data$nace))/100)

#OMITED DATA
lagged_variables <- c("failure", "iso", "control", "nace",
                      "shareholders_funds", "added_value",
                      "cash_flow", "ebitda", "fin_rev",
                      "liquidity_ratio", "total_assets",
                      "depr", "long_term_debt", "employees",
                      "materials", "loans", "wage_bill",
                      "tfp_acf", "fixed_assets", "tax",
                      "current_liabilities", "current_assets",
                      "fin_expenses", "int_paid",
                      "solvency_ratio", "net_income",
                      "revenue", "consdummy", "capital_intensity",
                      "fin_cons100", "inv", "ICR_failure",
                      "interest_diff", "NEG_VA", "real_SA",
                      "Z_score", "misallocated_fixed",
                      "profitability", "area", "dummy_patents",
                      "dummy_trademark","financial_sustainability",
                      "liquidity_return", "int_fixed_assets")
omitted <- na.omit(data[lagged_variables])
predictors <- c("control", "nace", "shareholders_funds",
                "added_value", "cash_flow", "ebitda",
                "fin_rev", "liquidity_ratio", "total_assets",
                "depr", "long_term_debt", "employees",
                "materials", "loans", "wage_bill", "tfp_acf",
                "fixed_assets", "tax", "current_liabilities",
                "current_assets", "fin_expenses", "int_paid",
                "solvency_ratio", "net_income", "revenue",
                "consdummy", "capital_intensity", "fin_cons100",
                "inv", "ICR_failure", "interest_diff", "NEG_VA",
                "real_SA", "misallocated_fixed", "profitability",
                "area", "dummy_patents", "dummy_trademark",
                "financial_sustainability", "liquidity_return",
                "int_fixed_assets")
formula <- as.formula(paste("as.factor(failure) ~",
                            paste(predictors, collapse="+")))

### Define samples
set.seed(123)
train_sample <- sample(seq_len(nrow(omitted)), size = nrow(omitted)*0.5)
train <- as.data.frame(omitted[train_sample,])
test <- as.data.frame(omitted[-train_sample,])
```

In this chunck we contruct the LOGIT model, we get the predicted probabilities and the confusion matrix.

We use as a threshold 0.3 (namely, $\hat{p}_i(Y_i = 1|X_i = x) > 0.3 \rightarrow \hat{y}_i = 1$), however this threshold can be moved up to 0.5 without any meaninful change.

```
log <- glm(formula, family = binomial, data = train)
prob_pred <- predict(log, type = 'response', newdata = test)
prediction <- as.numeric(prob_pred > 0.3) # change up to 0.5
cmlog=table(test$failure,prediction)
cmlog
```

```
##    prediction
##         0     1
##   0 43951   605
##   1  1381   473
```

In this chunck we use a LASSO model to select the variables with the highest predictive power for *false positives*.

```
test$prediction <- prediction
positive <- test[which(test$prediction==1),]
positive$iso <- as.numeric(as.factor(positive$iso))
positive$control <- as.numeric(as.factor(positive$control))
x <- as.matrix(as.data.frame(lapply(positive[predictors], as.numeric)))
y <- as.numeric(positive$prediction==1 & positive$failure==0)

mod <- cv.glmnet(x , y, alpha=1)
as.matrix(coef(mod, mod$lambda.1se))
```

```
##                                    1
## (Intercept)              0.495101483
## control                  0.084239097
## nace                     0.000000000
## shareholders_funds       0.000000000
## added_value              0.000000000
## cash_flow                0.000000000
## ebitda                   0.000000000
## fin_rev                  0.000000000
## liquidity_ratio          0.000000000
## total_assets             0.000000000
## depr                     0.000000000
## long_term_debt           0.000000000
## employees                0.000000000
## materials                0.000000000
## loans                    0.000000000
## wage_bill                0.000000000
## tfp_acf                  0.000000000
## fixed_assets             0.000000000
## tax                      0.000000000
## current_liabilities      0.000000000
## current_assets           0.000000000
## fin_expenses             0.000000000
## int_paid                 0.000000000
## solvency_ratio           0.001577534
## net_income               0.000000000
## revenue                  0.000000000
## consdummy                0.000000000
## capital_intensity        0.000000000
```

```
## fin_cons100               0.000000000
## inv                       0.000000000
## ICR_failure               0.000000000
## interest_diff             0.052601540
## NEG_VA                   -0.045676512
## real_SA                   0.000000000
## misallocated_fixed        0.000000000
## profitability            -0.005923519
## area                      0.005088934
## dummy_patents             0.000000000
## dummy_trademark           0.000000000
## financial_sustainability  0.000000000
## liquidity_return          0.436828513
## int_fixed_assets          0.000000000
```

```
row.names(as.matrix(coef(mod, mod$lambda.1se)))[which(as.matrix(coef(mod, mod$lambda.1se))!=0)]
```

```
## [1] "(Intercept)"      "control"          "solvency_ratio"
## [4] "interest_diff"    "NEG_VA"           "profitability"
## [7] "area"             "liquidity_return"
```

The indicator that seems to have the best disciminative power are: *control, solvency ratio, BID, negative AV, profitability, area, liquidity.* Hence, these indicators are the best to capture the component of "resistence" among highly distressed firms that makes them zombies.