Predicting French SME Failures: New Evidence from Machine Learning Techniques. *Christophe Schalck & Meryem Yankol-Schalk 2021*

Introduction

Small and Medium sized Enterprises (SMEs) play a crucial role in driving France's economic prosperity. In 2019, 99% of French Firms were SME's, accounting for 49% of French employees and contributing to 43% of the country's GDP. More recently, however, France's economy has been subject to volatility in SME failures which left the government with no choice but to introduce measures to reduce their financial constraints. This has shown a need for improved techniques to understand and predict SME failure.

The failure of an SME, as per Pretorius (2009), occurs when such a firm is no longer able "to attract new debt or equity funding to reverse decline" and is therefore unable to function. Failure is, therefore, the point at which the process of bankruptcy has reached the endpoint and operations cease. We consider a business with less than 250 employees and a turnover that does not surpass €50 million to be classed as SME – complying with EU Standards.

Numerous studies have investigated the financial variables associated with SME failure. Barboza et al., (2017) found that a firm's financial records, such as sales, operating margin, and solvency & liquidity, significantly improve the probability of SME failure in the United States. Further research has also confirmed the importance of financial variables in predicting business failure (Bunyaminu & Issah, 2012; Mselmi et al., 2017). Other key predictive variables include the size of the firm, with smaller establishments being the most vulnerable (Gupta. J et al., 2015; El Kalak. I & Hudson. R, 2016) and non-financial variables such as the economic conditions and CEO gender (Ptak-Chmielewska. A, 2019).

Traditional, empirical techniques that have been used to identify failure factors to aide predictions for SME failure often include statistical methods like discriminant analysis and logistic regression (Aziz and Dar 2006). However, in 2001 Shumway proposed the idea that such traditional techniques lack predictive power in default predictions, which catalysed a new branch of research investigating more sophisticated machine learning techniques to provide a better-performing alternative. Studies have found machine learning techniques

such as Artificial Neural Networks (ANNs) and Support Vector Machine (SVM) classification techniques to outperform traditional, statistical methods (Chaudhuri & De, 2011; Mselmi et al., 2017). The literature, therefore, suggests that there is room to improve on the traditional methods of forecasting SME failure, which must examine both financial and non-financial variables using machine learning techniques.

In their paper, Christophe and Meryem compare statistical and machine learning models: Dynamic Probit Model, Logistic Lasso Regression & Extreme Gradient Boosting, to find an accurate predictor of SME failure, and to find the variables that most effect the outcome.

Data

The dataset used by Christophe and Meryem was created from two sources. The first is the SIRENE system of INSEE, the French National Statistics Office, and the second is the 'Diane Database'. The INSEE identifies and registers every active French company and the SIRENE directory records their civil status, irrespective of their sector, legal form, or location. Altogether, the SIRENE directory holds records of 6.5 million companies. While the Diane Database includes a wealth of financial data for over 1.3 million French registered companies. Using these sources, this paper was able to amass a dataset of financial information of 579,892 SMEs from 2012 to 2018.

Table 1 shows the variables selected as the factors for predicting SME failure. The Score variable is a financial health indicator computed by the Diane software by finding the probability of bankruptcy based on the value of a z-score function which are based on financial ratios. Table 1 also shows the seven hypotheses formulated from the seven Factors that the predictive models aim to test (H1-7). Correlations between the variables were calculated, through Cramer's V based on the chi-squared statistic, to avoid bias in the predictive models. There were no strong associations between the variables.

Strategy

In this study, the failure prediction model was constructed as per Gupta et al., (2015) using a Cox proportional hazard model which allowed the effects of individual features of firm survival to be isolated. This paper is also novel in that it integrates text variables into its models using the TF-IDF approach from the field of Natural Language Processing (NLP). This

method finds the relative frequency of words in a document and then, to optimise their predictive models, the most repetitive words that did not portray useful information were excluded. This study compared a Dynamic Probit Model, Logistic Lasso Regression and XGBoost algorithm to find an accurate predictive model for French SME bankruptcy.

Table 1. Summary of data variables

Factor	Variable	Description	Type	Source	Associated Hypothesis
Finance	Revenue	Turnover (€)	Numerical	Diane	H1: Firm financial variables (revenue, sales, capital, and score) are negatively associated with SME failure risk
	Sales	Sales in units	Numerical	Diane	
	Capital	Equity (€)	Numerical	Diane	
	Score	Financial Health Indicator calculated by Diane Database (e.g., Risky, Safe)	Categorical	Diane	
Size	Employees	Number of Staff	Numerical	Diane	H2: SME size is negatively associated with SME failure risk
Non-Financial	Self-Ownership	O – No Employees N – With Employees S – Self Employee	Categorical	SIRENE	H3: Self Employment is negatively associated with SME failure risk
	Gender	Gender of the CEO (M/F)	Categorical	SIRENE	H4: The gender of the CEO is negatively associated with the SME failure risk
	Economic Conditions	Economic Indicator computed based standard deviation of growth rate relative to the median of each sector	Categorical	Diane	H5: Strong economic conditions are negatively associated with SME failure risk
Region	DOM-TOM	Region of HQ Location		Diane	H6: The regional location of an SME has a specific impact on its failure risk
Activity	Type of Activity	Business activity based on INSEE classification	Text	SIRENE	H7: The type of activity has a specific impact on SME failure risk

Dynamic Probit Model

Recent literature has integrated an autoregressive structure which extends the probit model, and the consequent dynamic probit model is expressed as:

$$P(y_t = 1|y_{t-1}, x_{t-1}) = \phi(\rho y_{t-1} + \beta x_{t-1} + \omega) \quad (1)$$

using maximum likelihood methods to estimate the parameters β (in vector form) by maximising the full sample log-likelihood function:

$$\boldsymbol{l}(\beta) = \sum_{t=1}^{T} (y_t \log\big(\phi(\pi_t) + (1 - y_t) \log\big(1 - \phi(\pi_t)\big)\big) \quad (2)$$

Logistic Lasso Regression

Logistic Regression is also a common, traditional technique used to estimate the probability of a binary response, determined by a sample's explanatory variables, x, the regression coefficients β (in column vector form):

$$P(y_i = 1) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \quad (3)$$

where the parameters are calculated by maximising the log-likelihood function:

$$I(\beta) = \sum_{i=1}^{n} (y_i x_i \beta - \log\big(1 + e^{x_i\beta}\big)) \quad (4)$$

However, when using multiple covariates, regressions often suffer from overfitting yielding a resulting model that is not optimal. An alternative extension is the use of Lasso Regression, which is a penalised regression (5) that maximises the log-likelihood function to find the regression coefficients (Park & Casella, 2008).

$$I_\lambda(\beta) = \sum_{i=1}^{n} (\, y_i x_i \beta - \log\big(1 + e^{x_i\beta}\big)) - \lambda \sum_{j=1}^{p} |\beta_j| \quad (5)$$

The values of the estimated coefficients produced by the regression depend on the selected values of λ. A sufficiently large λ allows the model to force some of the coefficients close to zero, to equal exactly zero (James et al., 2013). Lasso Regression also handles issues of multicollinearity and has the properties to minimise overfitting effects to numerical instability.

Extreme Gradient Boosting Algorithm (XGBoost)

Extreme Gradient Boosting is a predictive method which sequentially builds subtrees from an original tree whereby each subsequent tree reduces the errors of its predecessor. XGBoost has been gaining popularity due to its flexibility to function in a variety of settings and produce robust results. XGBoost uses a sample of multiple features x_i to predict the target y_i and finds the values of the function's parameters which best fit the training data and output values. In a data set of n observations and K additive functions, $\widehat{y_i}$ can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{6}$$

where f_k represents an independent tree and $f_k(x_i)$ is the predicted score given by the i-th sample of the k-th tree.

XGBoost approximates the function f_k by minimising the cost function $\mathcal{L}$ which is comprised of the loss function l and the regularisation term Ω – which penalises the complexity of the model and acts to omit the risk of overfitting. The loss function acts to measure the difference between the predicted $\hat{y}_i$ and the actual value y_i. The cost function is defined by:

$$\mathcal{L} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

here the regularisation term is defined as:

$$\Omega(f) = yT + \frac{1}{2}\lambda \sum_{j=1}^{T} w_{j_j}^2 \tag{8}$$

where y is the threshold parameter, w is the vector of scores on each leaf, λ is the regularisation parameter weights and T is the number of leaves. The optimal weights w_j^* of leaf j and its corresponding optimal value can be found using:

$$W_j^* = \frac{-\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{9}$$

$$\mathcal{L}^* = \frac{-1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + yT \tag{10}$$

where g_i and h_i are the first and second gradient orders of the loss function. The function of y is to act as the threshold for a gain, when the loss reduction is reduced to a value below y, algorithm function ceases. This reduces the intricacy of the model and as such the computational load.

Evaluation Metrics

They assess the model's ability to predict SME failure using the C-statistic (Harrell et al., 1982). Letting Z be the time of observation, η is the probable risk score based on a covariate vector, and a variable b such that $b = 1$ if the firm is in bankruptcy and $b = 0$ otherwise. The C-statistics are expressed as:

$$C = \frac{\sum_{i \neq j} 1\{\eta_i < \eta_j\} 1\{Z_i < Z_j\} b_j}{\sum_{i \neq j} 1\{Z_i < Z_j\} b_j} \quad (11)$$

for each pair of firms i and j. A model with a C value less than 0.5 can be considered poor, a value of 0.5 suggests the model's predictive power is no greater than random chance, and a value exceeding 0.7 implies the model has good predictive power, with 1 meaning it is perfect.

Another metric used to assess a predictive model's performance is the standard deviation of the residuals, calculated by the root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{2}\sum_{i=1}^{n} e_i^2} \quad (12)$$

here n is the number of observations and the difference between an observation and its corresponding prediction is denoted as e_i.

Results

The model's predictive abilities were modelled with the evaluation metrics as seen in Table 2. We can clearly see that XGBoost was the best-performing model with the highest c-statistic and lowest RMSE.

Table 2. Summary of the three model's evaluation metrics

Model	Variables	C-Statistic	RMSE
Dynamic Probit	Numerical Only	0.64	18
Lasso Logistic Regression	Numerical and Text	0.57	19
XGBoost	Numerical and Text	0.77	18

To then interpret the results of the XGBoost algorithm, this study used the Shapley (SHAP) value method which is a feature attribution method. This works by allocating a value to each feature for a prediction and, in doing so, ensures that there is feature consistency and stability (Meng, Y. et al., 2021) to provide consistnet and accurate results. Table 3 displays the impact of the model's SHAP values on the seven hypotheses proposed earlier.

Table 3. Summary of the status of the studies' hypotheses following model interpretation

Hypothesis	Status	Comments
H1: Firm financial variables (revenue, sales, capital, and score) are negatively associated with SME failure risk	Confirmed	Financial variables have a negative impact on failure prediction. Negative average SHAP (naSHAP) values of 0.282 for capital, 0.225 for revenue and 0.015 for sales
H2: SME size is negatively associated with SME failure risk	Confirmed	Number of employees has a naSHAP value of 0.053 for failure prediction
H3: Self Employment is negatively associated with SME failure risk	Confirmed	Ranking the SHAP values and their correlations to failure prediction ranked self-employment as the most significant factor forcing the model to predict SME failure
H4: The gender of the CEO is negatively associated with the SME failure risk	Rejected	No difference in the performance of female and male-owned new ventures
H5: Strong economic conditions are negatively associated with SME failure risk	Confirmed	Strong economic conditions are negatively correlated with failure prediction, aSHAP value of 0.032. Weak economic conditions are positively correlated, aSHAP value of 0.018
H6: The regional location of an SME has a specific impact on its failure risk	Confirmed	Running a business in Paris is a failure factor, with a naSHAP of 0.072. Likely due to strong competition. Localisation in the Northwest was also a failure factor, naSHAP of 0.035. Likely due to decreased geographical interdependence and economic agents
H7: The type of activity has a specific impact on SME failure risk	Confirmed	Running a holding company or a trade retail business had a negative impact on failure prediction (aSHAP value of 0.034 and .011 respectively). Fast food and real estate businesses had positive impacts on failure prediction (aSHAP values of 0.011 and 0.010 respectively)

Conclusion

This study found that, when compared to a Dynamic Probit and Logistic Lasso Regression, the XGBoost algorithm was the best model at predicting French SME failure. This aligns with the field's literature which suggests that sophisticated machine learning techniques can be more accurate and efficient predictive models. However, techniques such as Artificial Neural Networks (ANNs) and Support Vector Machine (SVM) classification methods have also been shown to produce powerful predictive models. Further research should be done to apply the problem of predicting French SME failure, and SME failure in other countries to more machine learning methods. It is also important to note that they were unable to implement the text variable into the dynamic probit model, perhaps diminishing its comparative power in predicting SME failure.

From the results produced by the XGBoost model, this study also confirmed the importance of several variables to SME failure. Self-ownership was the most significant predictor for

SME failure, with the size of the business and location also being strong predictors. The paper's findings show the key factors that managers of French SMEs should be examining to implement effective growth strategies and highlight potential warning signals to avoid business failure. The results are also useful to financial institutions through the identification of the main risk factors that lead to failure, thereby facilitating accurate evaluations of firm performance and risk before investment. Policy makers could also make use of the results as aides for improving support policies – especially focussing on supporting self-owning SMEs.

References

Aziz, M. & Dar, H. (2006). Predicting corporate bankruptcy. *Corporate Governance International Journal of Business in Society*. 6: 18-33.

Barboza, F. Kimura H., Altman, E. (2017). Machine leaning models and bankruptcy prediction. *Expert Systems with Applications*. 83: 405-417.

Bunyaminu, A., Issah, M. (2012). Predicting corporate failure of UK's listed companies: Comparing multiple discriminant analysis and logistic regression. *Journal of Finance and Economics*. 94: 6-22.
Chadhuri, A., De, K. (2011). Fuzzy support vector machine for bankruptcy prediction. *Applied Soft Computing*. 11(2): 2472-2486

El Kalak, I. and Hudson, R . (2016). The effect of size on the failure probabilities of SMEs: An empirical study on the US market using discrete hazard model. *International Review of Financial Analysis*. 43: 135–145.

Gupta, J., Gregoriou, A., and Healy, J. (2015). Forecasting bankruptcy for SMEs using hazard function: To what extent does size matter?. *Review of Quantitative Finance and Accounting*. 45(4): 845–869
Harrell, F., Califf, R., Lee, D., Rosati, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*. 247: 2543-2546

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Meng, Y., Yang, N., Qian, Z., Zhang, G.(2021). What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values. *J. Theor. Appl. Electron. Commer. 16*: 466-490.

Mselmi, N., Lahiani, A., and Hamza, T. (2017). Financial distress prediction: The case of French small and medium sized firms. *International Review of Financial Analysis*. 50: 67–80.
Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*. 103(482): 681–686.

Pretorius, M. (2009). Defining business decline, failure, and turnaround: a content analysis. *South African Journal of Entrepreneurship and Small Business Management*. 2(1): 1– 16

Ptak-Chmielewska, A. (2019). Predicting Micro-Enterprise Failures Using Data Mining Techniques. *Journal of Risk and Financial Management*. 12(30): 1–17.

Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business*. 74:101–124.