Universitat Autònoma de Barcelona

Facultat d'Economia i Empresa UAB

Bachelor's Degree Final Project

Faculty of Economics and Business

Title: Analyzing the Endogenous and Exogenous Causes of Firm Creation/Destruction Using Machine Learning.
Author: Valentino Agazzi Mandozzi
Tutor: Antoine Zerbini

Date: 08/05/2025

Grade: ________

Word count: 10235

Index:

1. Introduction

The objective of this project is to study the elements that have a direct influence on the creation/destruction of companies. I consider that the best way to analyze this kind of events is to consider separately the endogenous and exogenous causes of firm creation/destruction. To clarify, endogenous causes are those that exist because of the firm individual performance, decisions, behavior and activity, while the exogenous causes are those that do not directly depend on how the firm does, but in the environment in which it exists. More detailed explanations on the data and the features used to predict this phenomenon will be given later.

In order to understand this complex phenomenon, I will make use of a great tool that improves our analytic capabilities, and it is Extreme Gradient Boosting Regressor, an advanced Machine Learning tool used mainly in supervised learning tasks (such as this project) that is capable of detecting and understanding non-linear correlations and is robust to most of the usual issues, like missing values, outliers, heterogeneous scales in our data, and so on, that might damage the statistical analysis carried out by traditional tools or models, such as linear or logistic regression.

This analysis will allow us to mine meaningful insights that are going to be useful for us to understand better the nature of the creation and destruction process of firms and the implications that the business management and the local policies and other externalities have. The main idea is to generate valuable knowledge that can be useful for policymakers and their decision-making process in such an important topic.

The motivation behind this project is that personally, I consider that the firm creation/destruction index is such an underrated indicator when measuring the economic performance of a country. At the end of the day, companies are the main engine of an economy, are that element generating value at the time that they provide goods and services that will be consumed by the workers whose salary they pay. The number of firms in an economy has many implications, from the fact that the demand of labor is the element determining the level of income of the people while the supply of products is going to set the market prices at which these are sold. These are just a few examples; however, it is enough to remark that the importance of firm development and expansion are the base for sustainable growth in any economy.

Moreover, what I think I can add to the literature surrounding this topic, is that, as far as I was able to learn, most part of the studies and research programs carried out were developed using more basic statistical algorithms such as linear regression or random forest machines. Whether it is true that these approaches might be useful for obtaining meaningful insights and statistics, they are not considered "the state of art" of machine learning and predictive statistics such as other algorithms like XGBoost or Neural Networks might be, since these have a lot of handicaps such as missing data handling, incapability of detecting non-linear effects and correlations over the target variables, autocorrelation, heteroscedasticity, and so on. That's why I consider that developing this research with such a robust algorithm as Extreme Gradient Boosting is, will be very beneficial.

2. Analysis of Previous Literature

Starting with the research field surrounding the endogenous cases for firm bankruptcy, we can citate a variety of authors such as Segovia Vargas and Camacho Miñano (2018). Summarizing, they examine corporate viability in Spain's pre-bankruptcy proceedings with a focus on financial ratios and artificial intelligence diagnostics (support vector machines) in their research of bankruptcy within the finance academic field. This research studied 1,713 Spanish companies that included bankrupt firms together with healthy organizations to demonstrate that financial viability ratios as well as working capital ratios serve as key success factors for reorganization. The authors applied the C4.5 decision tree algorithm to achieve better than 97% accurate identifications of solvent and insolvent firms thus demonstrating the success of machine learning models. The study emphasizes the financial independence ratio of 0.33 because all companies beneath this point experience increased bankruptcy risks but those above face potential pre-bankruptcy restructuring. They concluded that AI-based methods should be included within legal and financial systems to improve decision-making abilities of judges and creditors and insolvency administrators which helps minimize both systemic failures and financial spill-over effects.

Continuing with the study of the endogenous causes, the researchers at Kadarningsih et al. (2021) directed their investigation toward Indonesian manufacturing companies to study how profitability creates different pathways to bankruptcy prediction. The authors validate their findings by using path analysis and Sobel tests which demonstrate that financial distress results from liquidity and leverage and operating capacity through their influence on profitability levels. The positive bankruptcy risk relationship with leverage closely follows conventional debt-risk rules yet operational execution and liquidity effects come exclusively from profitability outcomes. The study conducted with Warp PLS 5.0 reveals that strong profitability indicators like ROA help protect companies from bankruptcy while they operate under high leverage levels because these firms excel at running their operations efficiently. The model supports Altman's established method but extends its examination through profit evaluation within both outcome orientation and intermediary status thus providing an adaptable perspective for markets displaying heightened debt vulnerability under macroeconomic turbulence.

Moreover, Adamko and Chutka (2020) evaluate Altman's Z-score against specific Czech and IN indexes to run a worldwide analysis of bankruptcy prediction systems. The authors demonstrate the conflict between universal predictive techniques and situation-specific evaluation methods. The core principle of Altman's model demands particular adjustments including EBIT/Total Assets and asset turnover for IN05 index-based evaluations of Central European enterprises. Springate and Fulmer constructed liquidity and cash flow evaluation factors for their models that work best in their targeted sector yet other business evaluators still utilize compatible metrics. The authors advocate for hybrid analytical approaches to enhance performance evaluations because regulatory disparities force adaptions between universal financial ratios and local measurement techniques. Among other examples the authors support predictive methods that integrate AI since they establish their point by following Segovia Vargas and Camacho Miñano who state predictive models must adapt through economic development.

Starting with the field of the exogenous causes, Klapper, Amit, and Guillén (2008) describe the World Bank Group Entrepreneurship Survey as a specific cross-country time-series database measuring total and newly registered businesses across 84 countries. The authors established a thorough method to unify regulations across countries while defining business startup under registration as companies capable of taking on liabilities and making purchases. The authors demonstrate that business entry rates and density rates show direct proportional relationships with economic indicators and financial development alongside legal quality and governance indicators as well as electronic registration procedures. Both business facility for registration and absence of political corruption continue to drive additional firm establishment in countries with better governance despite controlling for income. The study authors illustrate by using case study examples that concentrated institutional and tax reforms led to significant increases in new firm creation because efficient registries along with e-government platforms serve as vital policy development tools. The authors in Klapper et al. explain that entrepreneurial growth boosts conventional business expansion and lowers poverty levels and economic decline yet confirm that this phenomenon might originate from robust economic development producing independent entrepreneurial capabilities. The authors support tracking entrepreneurship vitality by sustaining firm establishment metrics to bolster policy definition progress.

To finish with the research over the analysis of the endogenous causes, Hundt and Sternberg (2016) utilize Global Entrepreneurship Monitor individual-level microdata to study nascent entrepreneurship across 15 European countries by adding regional and national context variables. Their research design measures entrepreneurship through two distinct phases which determine both potential and nascent stages of business creation while separating general ventures from growth-focused ambitious business ventures based on innovation measures. The study implements hierarchical linear models which analyzes the individual effects (gender, age, education, income and employment status) and the regional influences (GDP value and unemployment data) as well as national elements (market regulation and government size) on entrepreneurial participation through Giddens's structuration theory framework. The analysis yields four key findings: first, both regional and national environments exert statistically significant effects on entrepreneurial activity alongside individual characteristics; second, although personal resources carry the greatest overall weight, the direction and magnitude of these effects shift across phases of the entrepreneurial process; third, the impact of micro and macro determinants intensifies as entrepreneurs move from intention to action and as their ambitions grow; and fourth, cross-level interactions reveal that context can amplify or dampen the influence of individual traits—for instance, deregulated markets may constrain opportunity motives among potential entrepreneurs but interact differently for ambitious nascent founders. The understanding of entrepreneurship requires attention to its hierarchical space-time structure according to Hundt and Sternberg who support more detailed contextual modeling based on better spatial units and additional place clustering indicators for revealing the multiple-step firm development process.

Decision makers can direct their intervention efforts effectively because of risk thresholds created from enhanced diagnostic accuracy using AI decision trees and path analysis. Prospective organizations ought to stay away from applying one-use standardized generic models since their combination with traditional financial ratio data together with industrial and regional specifics creates enhanced decision support. The development of modern bankruptcy prediction methods resulted from connecting state-of-the-art methodologies to contextual comprehension thus creating useful academic tools as well as operational platforms.

The methods used in previous studies depend mainly on traditional statistical techniques including path analysis, Sobel tests and hierarchical linear models as well as support vector machines and decision trees which belong to basic machine learning frameworks. Whether it is true that these established techniques delivered useful findings and strong predictive powers, it is also remarkable that they fall short of present-day machine learning methods regarding optimization efficiency combined with complex non-linear interaction abilities. My research gains remarkable methodological progress through XGBoost implementation because it enables effective feature importance assessment and better prediction precision and performs superior unbalanced data management through gradient boosting optimization.

3. Analysis of the Data

3.1. Bankruptcy Data Overview

This research starts with a bankruptcy examination of firms through financial data from 1999 to 2009 in the widely recognized Taiwan Economic Journal dataset. The dataset stands as a reliable source for research because it has received a top usability score of 10 from Kaggle users. The scoring system uses three criteria of data completeness and reliability together with workability to confirm data standards meet professional research requirements.

Y, and X1 through X95 represent 96 financial indicators in the presented dataset for evaluating company performance. The data comprises financial indicators including Return on Assets (ROA) and operating margins with profitability ratios alongside the current ratio for liquidity dimensions together with debt-to-equity and asset and net profit growth rates for performance assessment. The dataset contains seven efficiency metrics that illustrate how companies utilize their assets and manage their inventory.

The "Liability-Assets Flag" (X85) indicates whether company assets fall short of liabilities because it identifies cases where debt exceeds assets which serves as a solid indicator of financial distress. Companies that posted two consecutive years of negative net income can be recognized through the "Net Income Flag" (X94). The financial risk flags enable speedy identification of businesses that might become bankrupt.

The dataset stands out because absence of missing values enables investigators to perform analyses with full confidence. The data collection process occurring under Taiwan Stock Exchange regulations enforced the real-world industry standards for determining bankruptcy. Researchers and professionals utilize the dataset quite frequently based on data available on its Kaggle page which has attracted over 427,000 views while recording 49,000 downloads.

Two important details emerge from the dataset evaluation. The dataset demonstrates an unbalanced class distribution with only limited bankruptcies present among the studied companies (the frequency at about 5% in similar datasets). Additional techniques will handle this imbalance later by performing resampling and adjusting model weights. The dataset contains correlated financial ratios such as different expressions of Return on Assets (X1, X2, X3) and alternate forms of Net Value per Share (X16, X17, X18). The selection of foremost significant features should replace redundant elements in the dataset.

If you require to check for all the variables included in the data set, without discriminating on which were used in our model, you can visit the second appendix where all of them are listed.

3.2. Firm Creation Data Overview

The second section of research analyses firm creation trends across various nations by examining World Bank macroeconomic databases spanning from 2005 to 2018. Several business environment indicators make up the contents of the comprehensive dataset. We are focusing on the new firm density per 100.000 working-age people since this directly reflects entrepreneurial actions and serves as the main dependent variable for analysis. The interpretive factors shaping the explanatory variables consist of four specific tax burden measures which include profit tax rates, labor taxes, total tax rates and yearly taxation requirements for companies. The included fiscal metrics evaluate both the motivation and disincentive effects of taxation policies for new business creation.

In addition to taxation metrics the data contains essential information about business expenses together with how various institutions perform. The database measures time required to register property alongside property-related crime losses and problems stemming from electricity issues. The analyzed variables enable researchers to measure business obstacles within various national settings.

The database integrates vital economic metrics which consist of yearly inflation data and interest rates in addition to yearly foreign investment flows. The factors at the macroeconomic level establish financial conditions which determine business creation decisions. Finally the statistics about new firm establishment provide an independent measure of entrepreneurial activities which enhances the data alongside the density ratios.

The World Bank data maintains strong credibility levels which make them suitable for policy-related and economic research. Through its panel design the dataset can support cross-national research along with time-based analysis of the data. User-controlled validation of the dataset is necessary since it includes real-world data which contains missing values and rare outliers that affect variables measuring inflation rates and tax burdens specifically during economic downturns.

This analysis finds strong value in the dataset because of its complete coverage of fiscal policy benchmarks and institutional data and economic indicators. Understanding the birth and success or failure of businesses requires a thorough examination of tax policies together with infrastructure quality assessments which forms a detailed picture of the overall business environment. The data set contributes practical significance to both bankruptcy tests at the firm level while keeping its relevance accessible to entrepreneurs alongside policymakers.

4. Hypothesis

Our analysis on the endogenous causes of firm bankruptcy and the exogenous sources of firm creation/destruction at a country level, as mentioned before, will follow a separated analysis that will help us draw conclusions at the end, therefore, continuing with this separated structure, so will do our hypothesis.

Firstly, the endogenous signs leading to firm-level bankruptcy risk include debt-to-equity imbalance and weak cash flow handling as well as ongoing operational performance shortcomings. These align with Modigliani-Miller's bankruptcy cost theory (financial causation) and the resource-based view (operational causation). Business performance indicators such as declining net value growth together with reduced R&D investment demonstrate deep financial troubles because they signal both poor solvency and the necessity of innovation to escape bankruptcy.

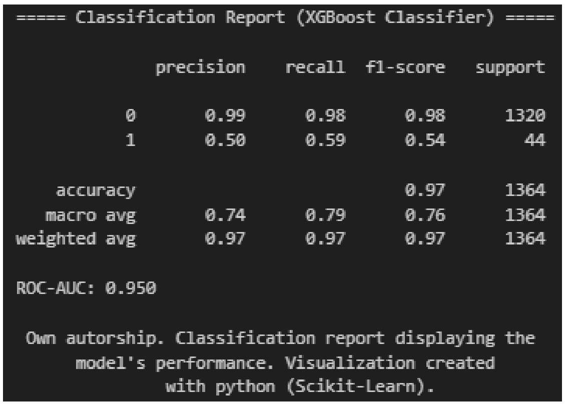
Secondly, I can advise that the relationship between entrepreneurial activity is negative to inefficient institutions combined with high government expenditures until reliable public infrastructure and incoming foreign national investments emerge. The market faces entry detentions from taxation disincentives meeting Laffer curve criteria and public procurement delays according to the institutional economics framework of North which combines with power infrastructure member issues that align with endogenous growth theory to heighten operational risks. Stable governance systems that function similarly to Europe along with capital inflows in accordance with Lucas' paradox about capital movement help regions neutralize these obstacles to build up entrepreneurial communities. Nonlinear policy success emerges from macro-level balance since governments must improve oversight functions in addition to basic finance systems and robust infrastructure.

5. Research Methods and Findings

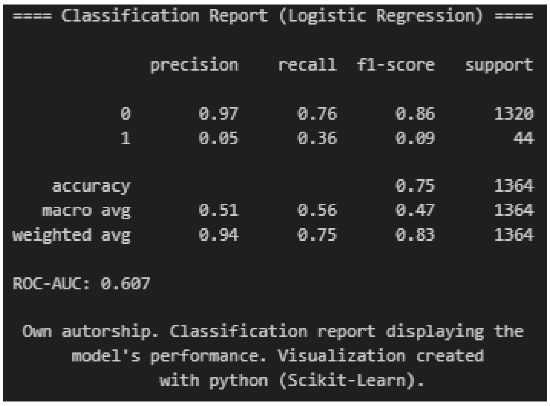
5.1. Bankruptcy Analysis

5.1.1. Model performance

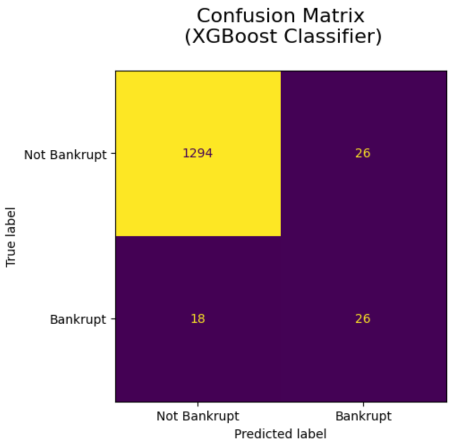
The model performance assessment begins with examination of the classification report shown below. Model performance assessment across multiple categories depends on classification reports and confusion matrices, that display essential metrics to determine success metrics and weak points. Considering class 0 as non-bankrupt observations and class 1 bankrupt as bankrupted firms, we can state that overall the model exhibits reliable total accuracy while also showing special patterns that require detailed analysis. For this case, I have also computed a Logistic Regression model using the same inputs, a classic tool for analysis used by economics and other sciences to study statistic correlations and make predictions. This model, instead, seems to have performed poorly, if we compare it with the XGBoost Classifier.



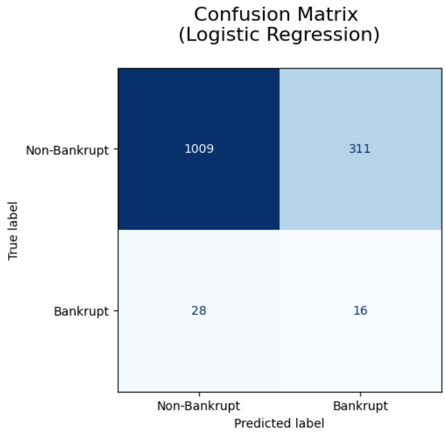
```
===== Classification Report (XGBoost Classifier) =====

              precision    recall  f1-score   support

           0       0.99      0.98      0.98      1320
           1       0.50      0.59      0.54        44

    accuracy                           0.97      1364
   macro avg       0.74      0.79      0.76      1364
weighted avg       0.97      0.97      0.97      1364

ROC-AUC: 0.950

 Own autorship. Classification report displaying the
      model's performance. Visualization created
              with python (Scikit-Learn).
```



```
==== Classification Report (Logistic Regression) ====

              precision    recall  f1-score   support

           0       0.97      0.76      0.86      1320
           1       0.05      0.36      0.09        44

    accuracy                           0.75      1364
   macro avg       0.51      0.56      0.47      1364
weighted avg       0.94      0.75      0.83      1364

ROC-AUC: 0.607

 Own autorship. Classification report displaying the
      model's performance. Visualization created
              with python (Scikit-Learn).
```



Own autorship. Confussion Matrix. Image generated using Python (matplotlib & scikit-learn).



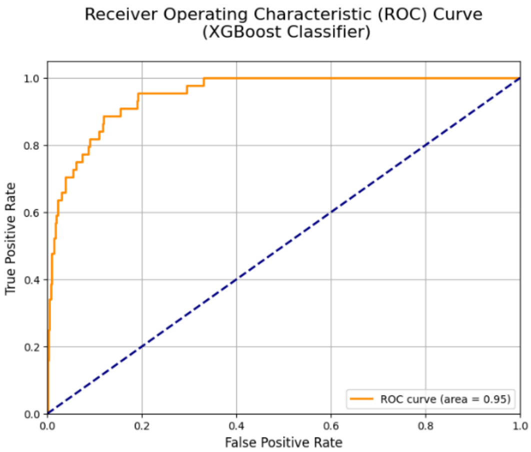
Own authorship. Confusion Matrix. Image generated using Python (matplotlib & scikit-learn).

Overall, we can say thaf for our XGBoost Classifier model, every metric shows outstanding results in class 0 at the beginning of our evaluation. The model demonstrates an accuracy rate of 0.99 when identifying class 0 predictions because 99% of them are correct. The model successfully recalls 0.98 instances from true class 0 samples. The F1-score at 0.98 indicates robust performance for the majority of the class and this score results from the precision and recall ratio combination through harmonic mean calculation. The good performance among the 1320 class 0 cases serves as the base upon which the model achieves its total accuracy rate of 0.97. The performance figures for class 1 demonstrate moderate results typical of imbalanced classification problems.

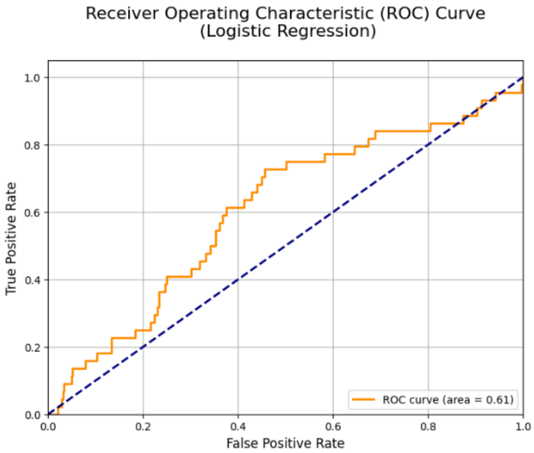
The accuracy rate for class 1 predictions stands at 0.50 while the model detects 59% of all actual class 1 instances. Such performance is reasonable in light of the 44 class 1 observations compared to the significantly larger number of 1,320 class 0 instances prevalent in unbalanced real-world datasets. The macro-level evaluation of 0.74 precision and 0.79 recall and 0.76 F1-score manifests the distinction between classes but does not acknowledge unequal class distributions. The weighted average totals 0.97 across all criteria proves the model excels at proper case management as it accurately handles most examples.

The logistic regression model identifies 97% of firm predictions as solvent but fails to detect 24% of the truly solvent firms. Despite reasonable performance, the computed F1-score at 0.86 demonstrates below benchmark results established by XGBoost which reached 0.98. Rephrase the following sentence. Present the information in direct language while keeping the text easy to understand along with verbalization when possible. The weakened model separability leads to overall accuracy declining to 0.75 and ROC-AUC dropping to 0.607. A high number of 311 firms from the non-bankrupt class were misinterpreted as bankrupt by logistic regression while it managed to correctly detect only 16 of the 44 actual bankruptcies because of its poor ability to handle extreme class imbalance. XGBoost exhibits a better capacity to detect the prevailing class and retrieve rare bankruptcy cases by performing near 0.97 in weighted averaging across performance metrics.

In contrast, regarding the ROC-AUC score value of 0.950 for our XGBoost model, and the visualized ROC Curve displayed below that offers essential additional information about the model's discriminatory power across all possible thresholds, we can state that an almost perfect ROC-AUC score as we are observing, represents excellent distribution separation between classes which shows that the model generates useful discriminative output probabilities. Even when some predictions are incorrect the model establishes relevant classification patterns according to its high AUC accuracy evaluation. Visual evaluation of the model's performance through the ROC curve presented in the attached chart shows excellent discrimination because it maintains a high true positive rate despite changes in false positive rates. Our XGBoost Classifier model demonstrates excellent sensitivity levels because its steep ascension occurs very near to the top-left corner of the chart.



Own autorship. ROC Curve. Image generated using Python (matplotlib & scikit-learn).



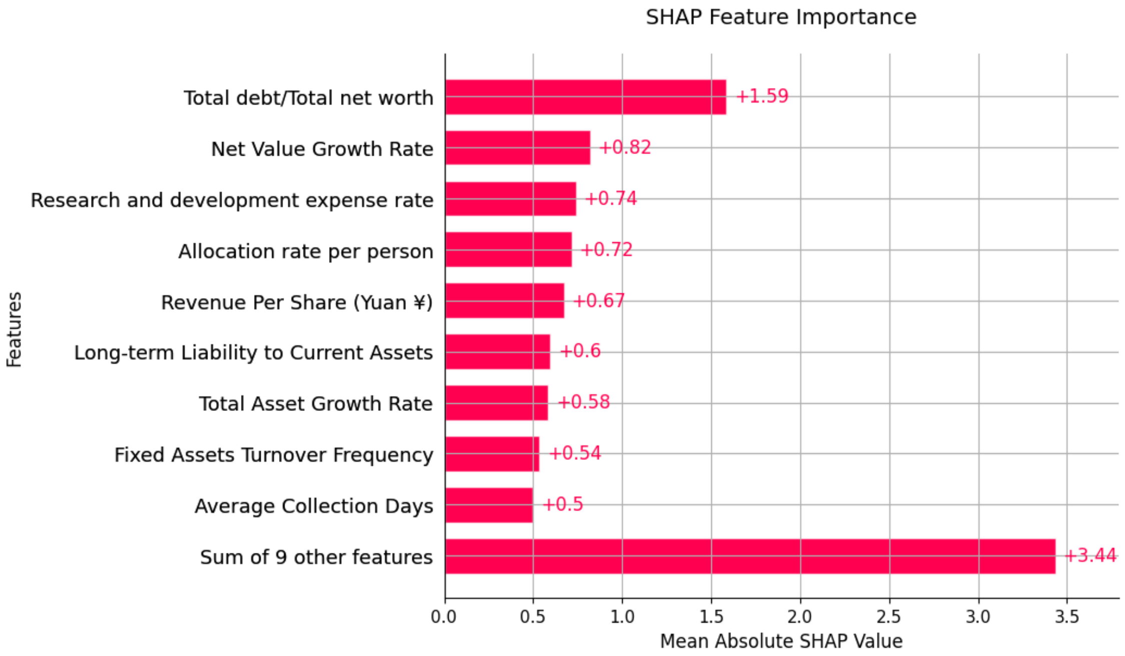
Own authorship. ROC Curve. Image generated using Python (matplotlib & scikit-learn).

The valuable aspect of our model derives from the integration of strong overall accuracy performance together with predictive capabilities that address both minority and majority classes. The model generates trustworthy baseline predictions as a result of its exceptional class 0 performance and delivers meaningful outcome predictions for class 1 cases although at an acceptable level. The sturdy ROC-AUC score verifies the model's base capability to separate classes so additional modification of thresholds or sampling methods may boost class 1 results without affecting excellent class 0 outcomes. The model shows practical value because it maintains a suitable ratio between precision and recall across different categories while demonstrating excellent discriminatory power as indicated by AUC.

5.1.2. Feature importance analysis

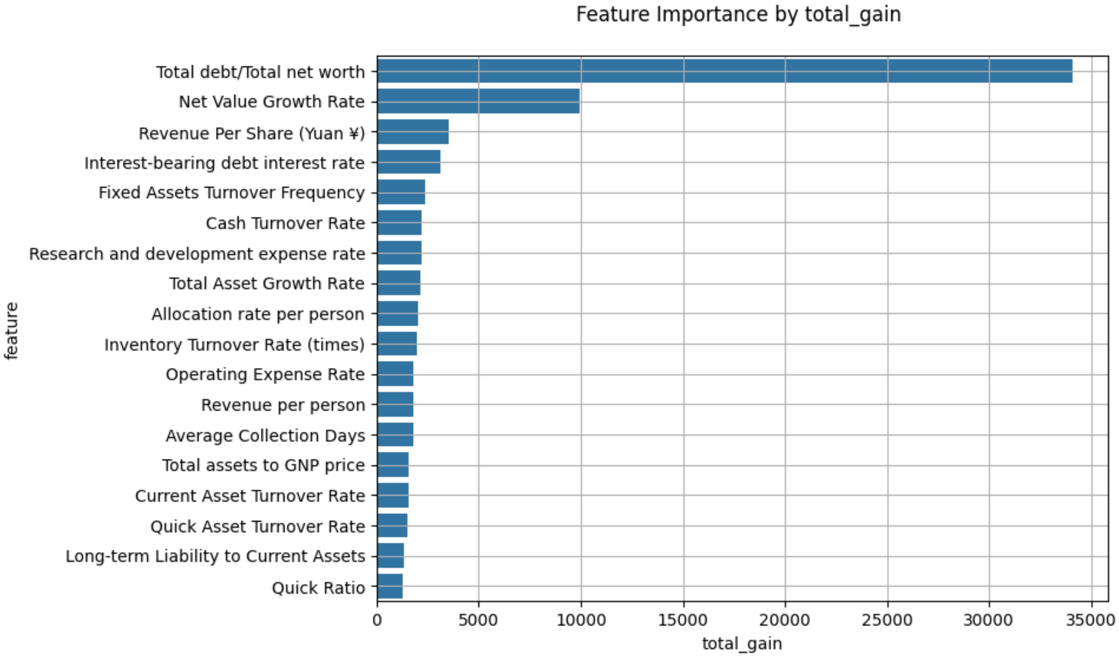
Two essential methods determine feature importance analysis within the XGBoost bankruptcy prediction model: SHAP (SHapley Additive exPlanations) values and XGBoost Gain metric. The two analytical methods demonstrate how financial ratios and variables affect bankruptcy prediction through complementary measurements of the model behavior.

Features get ranked through the SHAP summary plot based on their mean absolute prediction impacts computed as the average magnitude of SHAP values across observations. The influence rating comes from Total Debt/Total Net Worth since its mean SHAP value measures +1.51 which shows higher leverage ratios significantly boost predicted bankruptcy probabilities. Excessive debt compared to equity poses financial distress risks because it increases both fixed obligations while decreasing firm solvency in line with corporate finance theory. The second most vital element according to SHAP values (+0.82) reveals that organizations experiencing declining or negative equity growth face higher failure risk since sustainable capital expansion demonstrates importance. The two vital elements of operational efficiency and innovation capacity called Allocation Rate per Person and R&D Expense Rate support analysis of long-term viability potential. The more than three combined possible factors (+3.76) show bankruptcy risk results from interconnected financial, operational and macroeconomic variables.


SHAP Feature Importance

Own autorship. Barplot displaying feature importance by Shap.
Created using Python (matplotlib & seaborn)

The Total Gain metric determines feature value through loss reduction measurement that results from feature splits in predictive modeling. The primary place of bankruptcy prediction rests with Total Debt/Total Net Worth according to the Total Gain evaluation (34113.54). The findings align with trade-off theory from capital structure theories because they show that high leverage causes default risks to increase. The Net Value Growth Rate shows an overall profit increase of 9961.83 which strengthens proof for equity diminishment as a cause of financial problems. The features Revenue Per Share along with Interest-Bearing Debt Interest Rate demonstrate the significance of profitability and debt servicing costs as per liquidity constraint theories (Total Gain values of 3502.34 and 3150.92 respectively). Despite overall consistency with SHAP analysis the SHAP approach gives superior results to Total Gain analysis when it comes to ranking Long-Term Liability to Current Assets higher than Gain does. This indicates that SHAP handles features that show nonlinear behavior or context dependence more effectively through its local explanation methods.



Own autorship. Barplot displaying feature importance by total_gain.
Created using Python (matplotlib & seaborn)

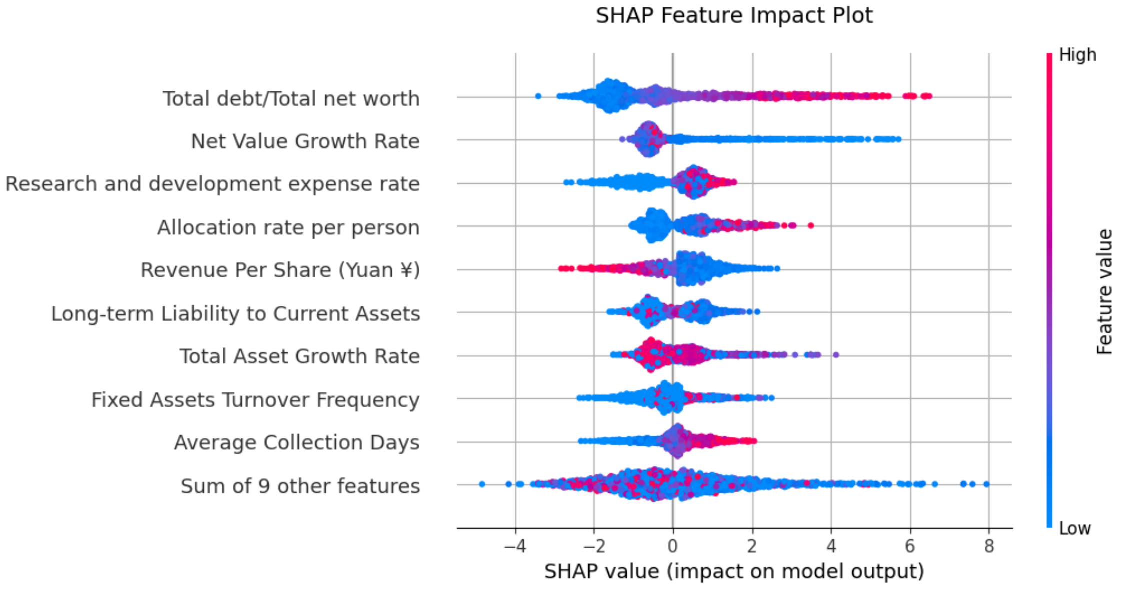
By using financial theory analysis we validate leverage and liquidity metrics because they meet the requirements of bankruptcy models such as Altman's Z-score and similar models that emphasize solvency and short-term debt. The interpretation of Net Value combined with Total Asset Growth rates supports the declining firm hypothesis because declining asset growth indicates diminishing market competitiveness. Company operations and cash management performance is reflected by efficiency ratios which focus on Fixed Assets Turnover and Average Collection Days because they ensure sustainable cash flows. R&D Expense Rate and Allocation Rate per Person enhance bankruptcy risk assessment because both factors indirectly affect risk exposure according to resource-based view theories. The combined findings from SHAP metrics and Gain analysis reveal that bankruptcy prediction happens through leverage and growth elements alongside liquidity attributes of the business while SHAP gives detailed direction and Gain demonstrates the predictive strength.

5.1.3. Understanding the impact of the features on the predictions

Our study uses two specification tools from SHAP which we apply to analyze feature effects on our bankruptcy prediction model. These tools include beeswarm plots and waterfall plots.

The beeswarm plot shows a universal assessment of feature influence over all cases from the dataset. A feature-specific vertical ordering determines stripe position in the graph which shows descending importance from top to bottom. Feature values on the x-axis show positive SHAP contributions toward bankruptcy and negative values toward non-bankruptcy condition predictions of the model. Each data point shows a solitary event which displays its normalized value as red for high (1) and blue for low (0). The points positioned on the x-axis indicate the scale and type of influence each feature provides to the model prediction. High values of features with red point clustering in the positive sector consistently raise the bankruptcy risk.

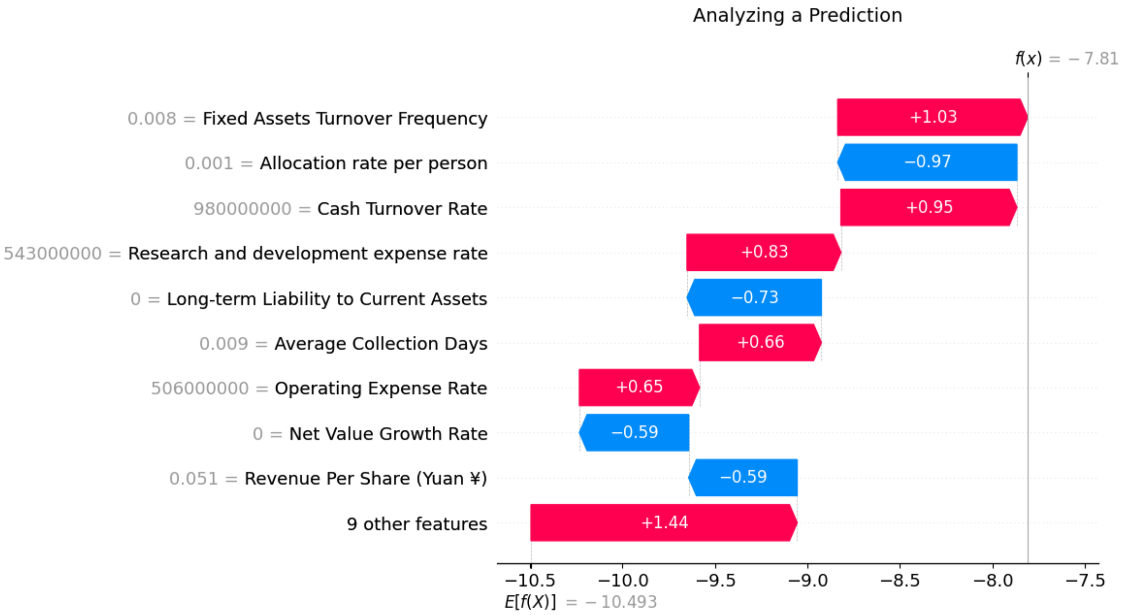
According to the beeswarm plot financial leverage metrics along with growth indicators drive most predictions from the model. Total Debt/Total Net Worth demonstrates the most significant impact because points with high values cluster toward the positive SHAP side which indicates better bankruptcy risk performance similar to how excessive debt decreases financial stability. The Net Value Growth Rate model performs differently because low growth rates make instances positive for bankruptcy prediction yet instances with high growth reduce risk. Statistics indicate that Research and Development Expense Rate and Allocation Rate Per Person produce small but dependable changes in the prediction of bankruptcy. Revenue Per Share functions as a two-part indicator because risk minimization happens when values remain moderate except when they reach extreme high or low points which leads to increased volatility according to SHAP values. The patterns indicate how the model depends on financial health and operational performance to determine bankruptcy risks of companies.



Own autorship. Beeswarm plot displaying feature impact on Predictions.
Created using Python (matplotlib & seaborn)

The waterfall plot examines how a single instance gets predicted through step-by-step analysis. Starting from the baseline expectation the model prediction begins at −10.493 (the average model output) and each subsequent row reveals the effect of feature values on the prediction output. In the left section of the table the instance features appear while SHAP contribution values appear to the right. The predictions toward bankruptcy receive upward pressure from blue bars and downward pressure from red bars. The overall prediction combines every contribution together. The plot demonstrates which characteristics were the key determinants for this particular case by helping explain their ranking significance.

A waterfall plot shows a specific case where localized feature interactions lead to prediction determination. The frequency of turning fixed assets represents 0.008 in the Fixed Assets Turnover Frequency (0.008) features that generate +1.03 SHAP units which indicates underutilized fixed assets as a risk factor. The market risk lowers by -0.97 units as a result of the Allocation Rate Per Person (0.001). When analyzing Research and Development Expense Rate (¥543M) as a factor the contribution was negative (-0.73) even though research intensity may deplete the cash balance. This company's lack of long-term liabilities at 0.0 for Long-Term Liability to Current Assets creates additional risk reduction of -0.59 units. The share value of revenue falls at ¥0.051 yet contributes minimally to +0.65 because the model treats debt and operational metrics as more important than revenue measurement in this case. The overall impact of these variables alters the prediction by +1.03 units from its initial position to produce a comprehensive bankruptcy risk assessment.



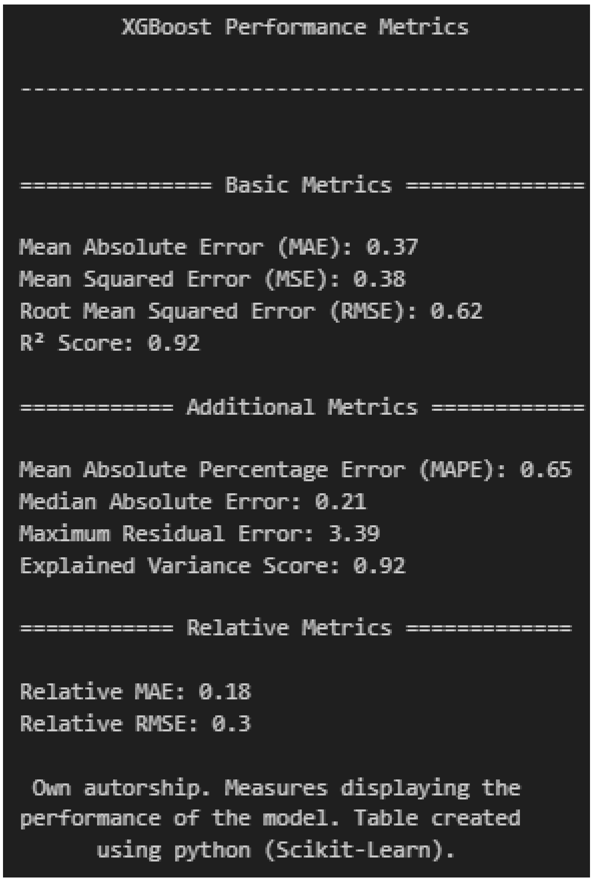
Own autorship. Waterfall plot showing how a prediction is reached.
Created using Python (matplotlib & seaborn)

Together, these plots validate the model's alignment with financial theory while exposing granular decision pathways. Globally, debt ratios and growth metrics are paramount, corroborating established bankruptcy predictors. By combining global and local perspectives, we confirm that the XGBoost framework not only identifies dominant risk factors but also adapts to idiosyncratic scenarios, enhancing its reliability for bankruptcy prediction.

5.2. Country level analysis

5.2.1. Analyzing the model's performance

In order to analyze the performance of the model, we can rely on the values displayed in the figure below. Here, the most important metrics used to analyze the efficiency and effectiveness of a ML regression model are depicted.



```
        XGBoost Performance Metrics

--------------------------------------------


=============== Basic Metrics ==============

Mean Absolute Error (MAE): 0.37
Mean Squared Error (MSE): 0.38
Root Mean Squared Error (RMSE): 0.62
R² Score: 0.92

============ Additional Metrics ============

Mean Absolute Percentage Error (MAPE): 0.65
Median Absolute Error: 0.21
Maximum Residual Error: 3.39
Explained Variance Score: 0.92

============ Relative Metrics =============

Relative MAE: 0.18
Relative RMSE: 0.3
```

Own autorship. Measures displaying the performance of the model. Table created using python (Scikit-Learn).

The predictive model demonstrates effective metric results which confirm its ability to analyze variables that impact new company density per 100,000 inhabitants. The model explains 92.1% of the target variable's variations through its predictor variables because its R^2 reaches 0.921. The model demonstrates excellent explanatory power because it effectively reproduces the fundamental data relationships.

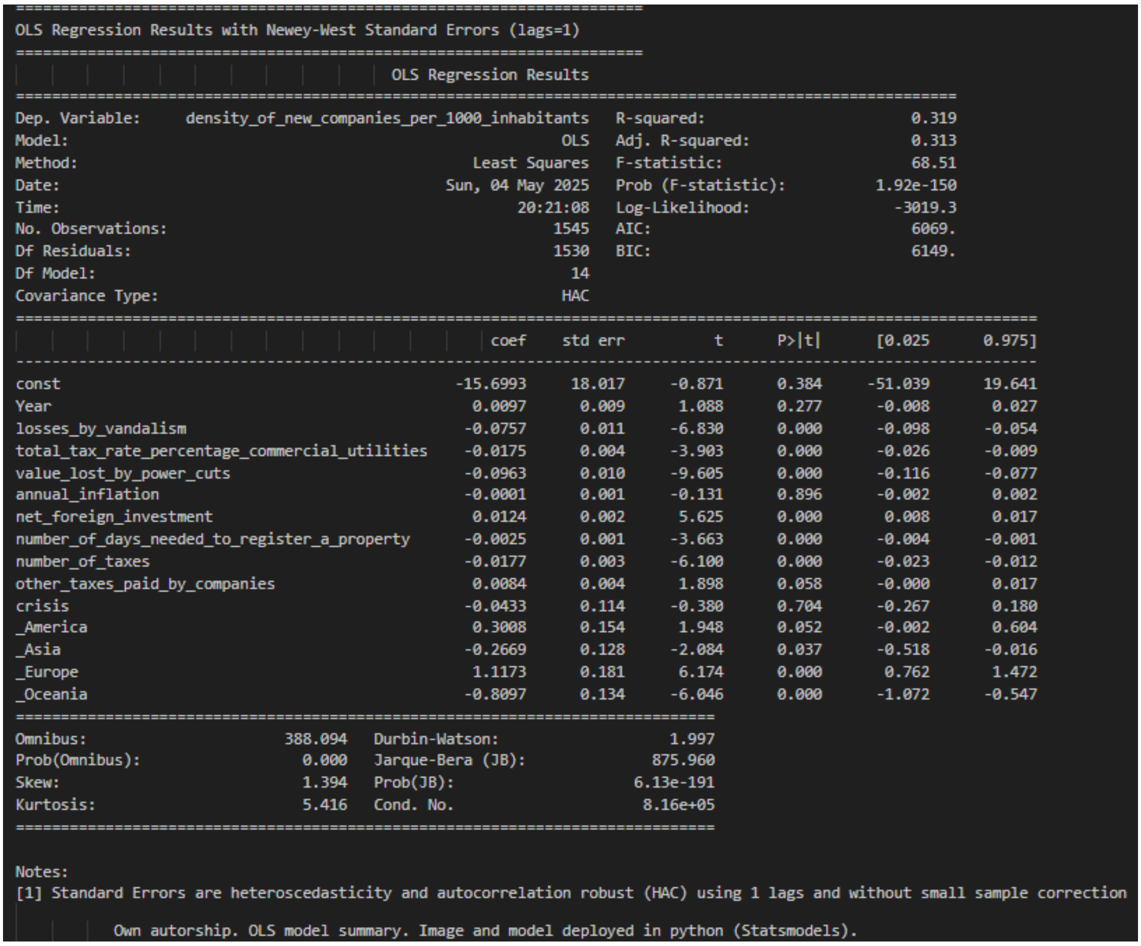
The Mean Absolute Error of 0.371 and Root Mean Squared Error of 0.616 indicate that predictions from the model show deviations below half a unit in the measurement of companies per 100,000 inhabitants on average. The Median

Absolute Error value of 0.205 represents the precise quality of the model since it demonstrates that half of the predictions fall within 0.205-unit errors. The Maximum Residual Error of 3.392 exists along with some difficult outliers but these cases remain exceptional due to the substantial difference between the median and mean errors.

A Mean Absolute Percentage Error (MAPE) of 64.8% indicates substantial relative error until we consider the target variable contains many small values. This makes percentages misleading about the largest absolute deviations. Statistical calculations using relative measures show that prediction errors along with target variable mean exhibit low discrepancy because Relative MAE stands at 18.0% and Relative RMSE reaches 29.8%.

A model Explained Variance Score of 0.921 matches closely with the statistical R2 score thus demonstrating the model can generate accurate predictions outside of training data. The multiple assessment metrics confirm that the model achieves reliable prediction of new observations in addition to producing accurate data fits thus making it a useful analytic instrument for policy decisions. All error measurements demonstrate consistency which confirms that the model effectively gauges practical and statistical elements of economic measures on entrepreneurial activity.

As well as we did previously, once again, a classic Linear Regression model using the same data we used for our XGBoost regression. Overall, observing the description of our classic Linear Regression model, the independent variables explains a low 32 % of the total dataset variation while demonstrating a significant F-statistic of 68.5 ($p \ll 0.001$) yet its residuals present substantial heteroskedasticity based on AIC scores of 6 069 and BIC scores of 6 149, issue that required to be solved using the Newey-West standard errors (nº of lags chose through PACF). Meanwhile, our XGBoost Regression, besides being robust to heteroscedasticity, reaches an R^2 of 0.92 thus accounting for more than nine-tenths of the outcome variability while generating low RMSE of 0.62 and MAE of 0.37—a fraction of traditional linear fit errors. The tree-based prediction model maintains all of its errors below the maximum level of 3.39 while producing only 0.21 median absolute errors compared to the linear model's much higher and more resistant residuals. An evaluation by relative MAE/RMSE of 0.18:0.30 together with the 0.65 % MAPE value proves that XGBoost generates both precise and uniform predictions throughout the complete sample. The linear regression model maintains a traditional single vertical regression line that fails to adjust to nonlinear patterns in the data thus producing large absolute and percentage-based prediction errors.

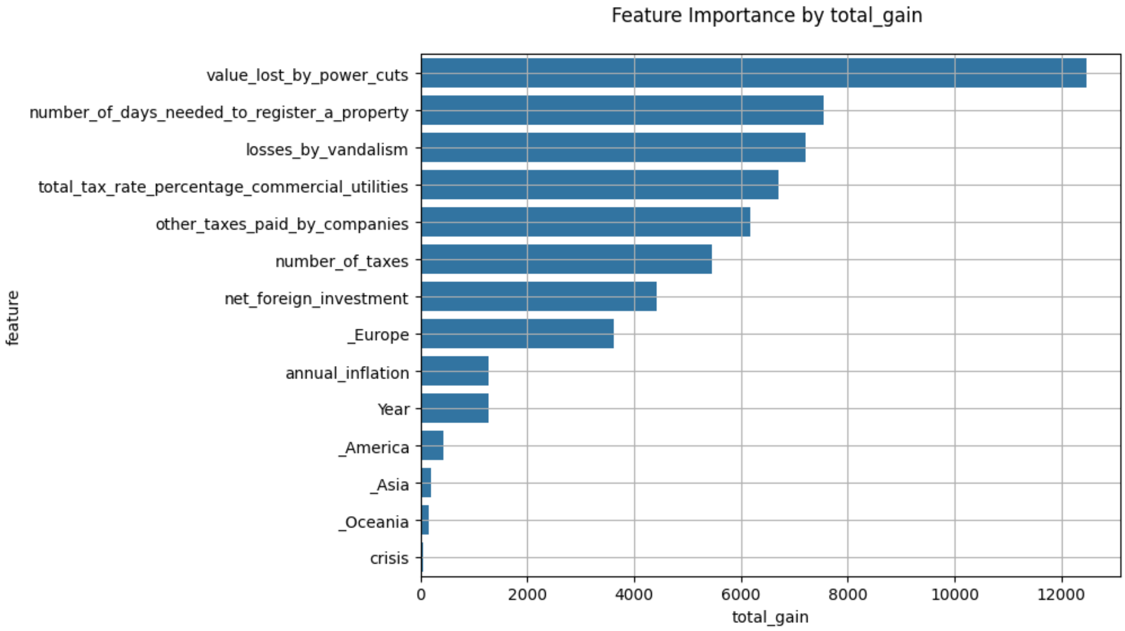


```
======================================================================
OLS Regression Results with Newey-West Standard Errors (lags=1)
======================================================================
                                          OLS Regression Results
=========================================================================================================
Dep. Variable:     density_of_new_companies_per_1000_inhabitants   R-squared:                       0.319
Model:                                                       OLS   Adj. R-squared:                  0.313
Method:                                            Least Squares   F-statistic:                     68.51
Date:                                           Sun, 04 May 2025   Prob (F-statistic):          1.92e-150
Time:                                                   20:21:08   Log-Likelihood:                -3019.3
No. Observations:                                           1545   AIC:                             6069.
Df Residuals:                                               1530   BIC:                             6149.
Df Model:                                                     14
Covariance Type:                                             HAC
==================================================================================================================
                                                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
const                                            -15.6993     18.017     -0.871      0.384     -51.039      19.641
Year                                               0.0097      0.009      1.088      0.277      -0.008       0.027
losses_by_vandalism                               -0.0757      0.011     -6.830      0.000      -0.098      -0.054
total_tax_rate_percentage_commercial_utilities    -0.0175      0.004     -3.903      0.000      -0.026      -0.009
value_lost_by_power_cuts                          -0.0963      0.010     -9.605      0.000      -0.116      -0.077
annual_inflation                                  -0.0001      0.001     -0.131      0.896      -0.002       0.002
net_foreign_investment                             0.0124      0.002      5.625      0.000       0.008       0.017
number_of_days_needed_to_register_a_property      -0.0025      0.001     -3.663      0.000      -0.004      -0.001
number_of_taxes                                   -0.0177      0.003     -6.100      0.000      -0.023      -0.012
other_taxes_paid_by_companies                      0.0084      0.004      1.898      0.058      -0.000       0.017
crisis                                            -0.0433      0.114     -0.380      0.704      -0.267       0.180
_America                                           0.3008      0.154      1.948      0.052      -0.002       0.604
_Asia                                             -0.2669      0.128     -2.084      0.037      -0.518      -0.016
_Europe                                            1.1173      0.181      6.174      0.000       0.762       1.472
_Oceania                                          -0.8097      0.134     -6.046      0.000      -1.072      -0.547
==============================================================================
Omnibus:                      388.094   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              875.960
Skew:                           1.394   Prob(JB):                    6.13e-191
Kurtosis:                       5.416   Cond. No.                     8.16e+05
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using 1 lags and without small sample correction

           Own autorship. OLS model summary. Image and model deployed in python (Statsmodels).
```

5.2.2. Feature importance analysis

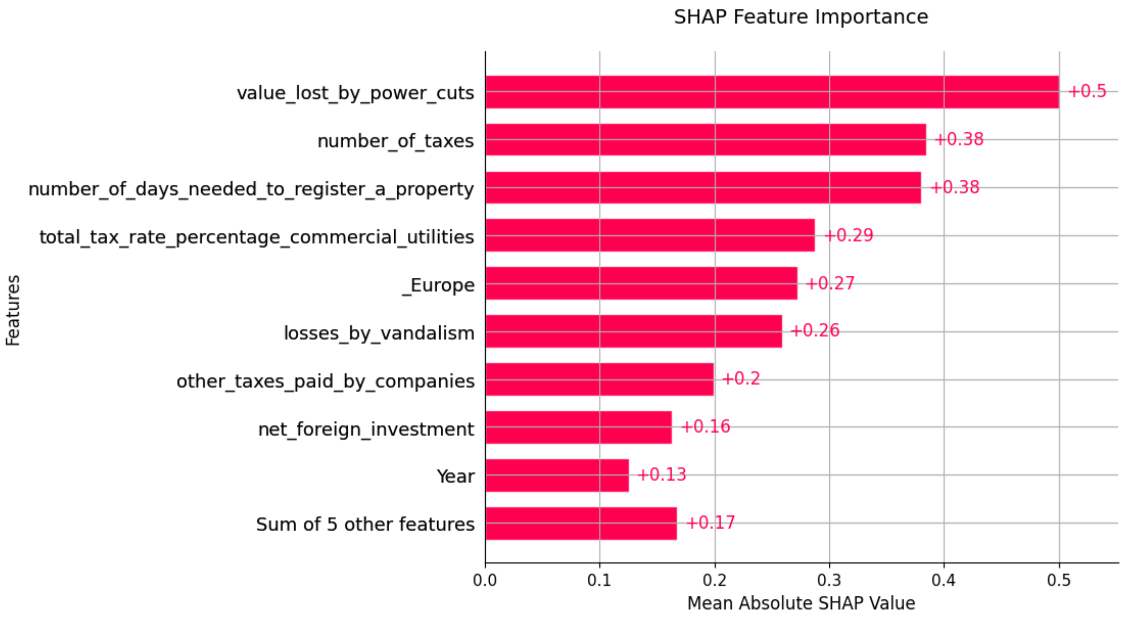
Evaluation of this model's feature importance uses the same approach which was employed earlier. The main source of truth derives from total gain and shap values.

A visual representation of total_gain reveals that value_lost_by_power_cuts stands out as the primary predictor with its total_gain value reaching close to 12,000 before any other variable in the second place. The result of unreliable infrastructure led to economic instability that causes business conditions to evolve negatively toward discouraging entrepreneurial activity. Bureaucratic inefficiency emerges as a top barrier against new business creation because the time needed for property registration (number_of_days_needed_to_register_a_property) totals approximately 10,000 in total_gain value. The business activity suffers due to high tax rates because operational costs rise from factors including total_tax_rate_percentage_commercial_utilities (~8,000) and number_of_taxes (~7,500). The two variables generate different effects in business performance as vandalism damages trust while foreign investment into local markets indicates the market's appeal to international investors. The sensitivities of entrepreneurship measurements between Europe and America and Asia fall within the range ~4000 to 3000 which indicates that economic and cultural factors shape entrepreneurial contexts. The data indicates that both annual_inflation (~2,500) and crisis (~1,500) discourage business commitments through their impact on macroeconomic stability.



Own autorship. Barplot displaying feature importance by total_gain.
Created using Python (matplotlib & seaborn)

The SHAP analysis verifies these key findings by providing additional direction to the results. value_lost_by_power_cuts emerges as the variable with the largest SHAP value of 0.5 even though its positive direction indicates that severe power cuts lead to less operating businesses since disruptive infrastructure failures reduce business sustainability rates. The negative SHAP value of -0.38 confirms that increasing numbers of taxes negatively affect entrepreneurship. Enterprise density increases when the number of days needed to register properties decreases (SHAP +0.38) which demonstrates the importance of cutting down regulatory process times. The regional observation of Total_tax_rate_percentage_commercial_utilities (-0.29) along with Europe (+0.27) demonstrates that utility tax rates negatively affect firm establishment and Europe’s positive influence stems from its stable institutional framework. Security and investment variables appear prominently in SHAP values because vandalism results in lower confidence (+0.26) and foreign capital injection (+0.2) drives economic growth. The analytical method known as SHAP value (Sum of 5 other features) shows that various seldom considered variables in combination lead to prediction outcomes thus demonstrating the model's comprehensive understanding of economic complexities.

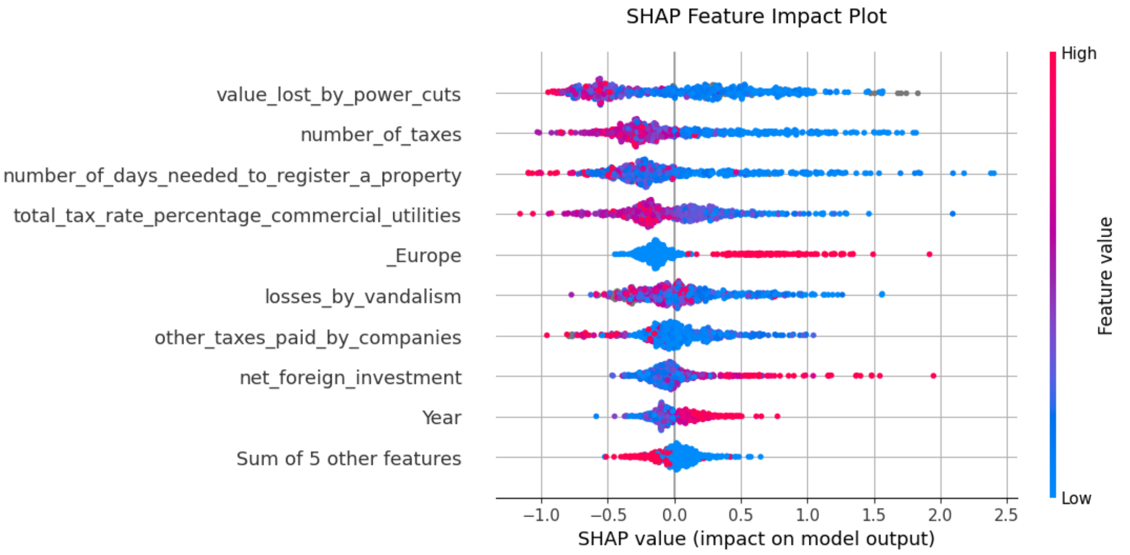


Own autorship. Barplot displaying feature importance by Shap.
Created using Python (Matplotlib, Seaborn & Shap)

5.2.3. Understanding the impact of the features on the predictions

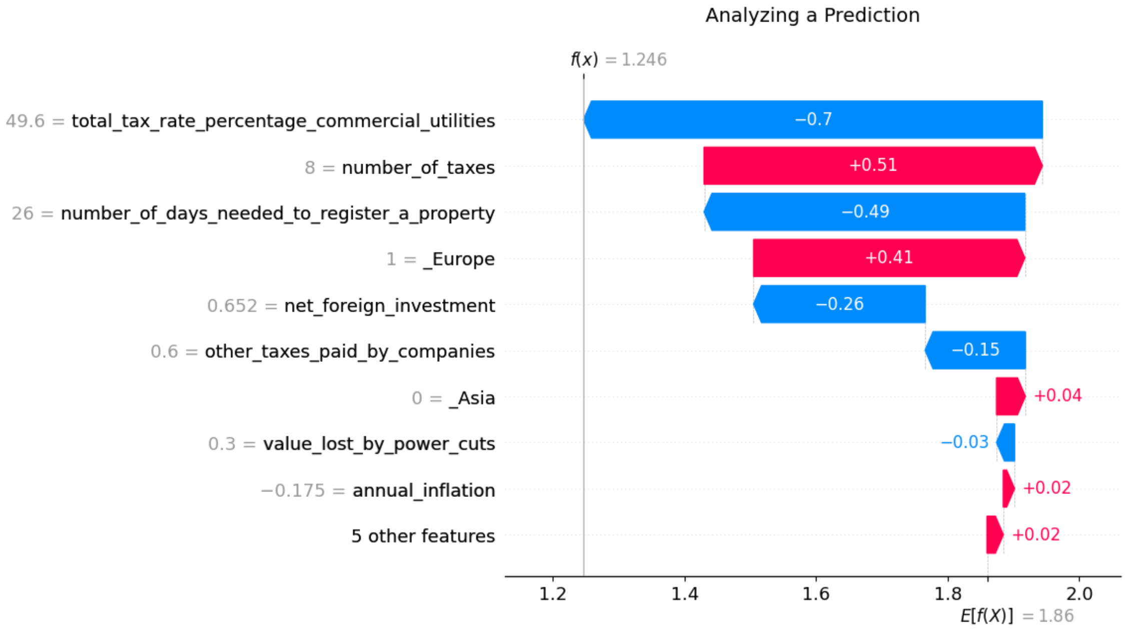
Continuing with our analysis. We will make use once again of the beeswarm and the waterfall plot in order to understand how the model's features affect our target variable as they vary.

In this case, the beeswarm plot shows that regulatory and macroeconomic elements hold essential importance for establishing new businesses. Total Tax Rate Percentage Commercial Utilities proves itself as the leading factor which negatively affects firm creation based on high-tax rate distribution (red points on the negative SHAP side). Data shows that higher investment levels from foreign capital (red points on positive side) correspond with increased firm creation since foreign capital establishes a positive relationship between the two variables. The duration it takes to register properties (red points) plays an important role in limiting new firm establishment because long processing times indicate inefficient bureaucratic administration. The effects from Europe as a region come out differently across the data points indicating that regional elements function together with other relevant factors. The spread of Value Lost by Power Cuts shows minimal variation which indicates its negative effects occur steadily.



Own autorship. Beeswarm plot displaying feature impact on Predictions.
Created using Python (matplotlib & seaborn)

Local dynamics become visible in representative case waterfall plots because they affect prediction results. The contribution of Total Tax Rate Percentage Commercial Utilities (49.6%) to firm creation is measured at -0.7 SHAP units indicating a negative effect on firm conception. External capital through Net Foreign Investment helps balance adverse effects from high-tax environments since it brings +0.51 units to the prediction. Lastly the prediction experiences a -0.49 reduction due to the geographic variable being set to Europe which suggests European markets either show market saturation or intense business competition. According to the global data, Number of Taxes (8) shows a medium negative effect (-0.26) and Number of Days Needed to Register a Property (26 days) creates an additional -0.15 effect on entrepreneurship. Although Value Lost by Power Cuts amounts to 0.3 the small negative impact it produces is maintained (-0.03) demonstrating how minor infrastructure problems build up. These multiple factors together decrease the model prediction by 0.614 units thus demonstrating how model priorities different variables to deliver balanced results.



Own autorship. Waterfall plot showing how a prediction is reached.
Created using Python (matplotlib & seaborn)

The various business plots demonstrate the different elements which propel new firm formation. Statistical data confirms that tax burdens and administrative efficiency drive firm creation worldwide based on existing economic research. The waterfall representation demonstrates that regional and contextual elements including foreign investment amounts and geographic position adjust these findings. High taxes always present universal challenges yet strong foreign capital inflow reduces their adverse consequences. New firm creation faces limitations in Europe’s developed markets, but additional variables including investment and infrastructure impact this restriction.

6. Conclusions

The development of my thesis relied on an investigation of internal and external elements that trigger firm start-ups and closures that utilized Extreme Gradient Boosting to assess both detailed and broad-scale data sets. My research employed XGBoost to understand variable relationships and deal with disproportioned data points beyond traditional statistical methods that investigated firm birth and death patterns.

The combination of high leverage ratios particularly Total Debt to Total Net Worth as well as declining equity growth together with weak operational measures expressed through R&D Expense Rates and Allocation Rate per Person led to bankruptcy in firms. New business formation density faced negative effects from unstable power infrastructure that caused revenue loss through power outages along with high commercial utility tax rates which combined with long administrative procedures at the country level but foreign direct investment along with developed institution frameworks supported entrepreneurial activity. This research proves the necessity of corporate financial adjustments by companies along with simpler regulations and reduced taxes for governments to reduce barriers to business startups. The combination of economic measures establishes resilience for the economy along with safe business turnover preservation.

The research study encounters numerous constraints because it depends on bankruptcy reports from the Taiwan Economic Journal from the 1990s and uses World Bank panel data that concludes at 2018 without considering COVID-19 impacts. The analysis of this framework should examine two recommended methods by using recent data after 2018 together with micro-level industrial information and executing neural network technologies to forecast time series while performing policy distinction through causality analysis. Research that combines survival behavior analysis across firm levels with regional survival patterns will develop insights between economic elements extending from regions to the national scale.

7. Appendix.

I. Source Code

The code used to develop this study can be found in my GitHub page. It was divided into two parts, the first one, containing the firm-level analysis, and the second one, containing the country level analysis.

- Find the firm-level bankruptcy analysis code here.
- Find the country level firm creation analysis here.

II. All the variables from the Company Bankruptcy Prediction Dataset (used and dropped).

Y - Bankrupt?: Class label

X1 - ROA(C) before interest and depreciation before interest: Return On Total Assets(C)

X2 - ROA(A) before interest and % after tax: Return On Total Assets(A)

X3 - ROA(B) before interest and depreciation after tax: Return On Total Assets(B)

X4 - Operating Gross Margin: Gross Profit/Net Sales

X5 - Realized Sales Gross Margin: Realized Gross Profit/Net Sales

X6 - Operating Profit Rate: Operating Income/Net Sales

X7 - Pre-tax net Interest Rate: Pre-Tax Income/Net Sales

X8 - After-tax net Interest Rate: Net Income/Net Sales

X9 - Non-industry income and expenditure/revenue: Net Non-operating Income Ratio

X10 - Continuous interest rate (after tax): Net Income-Exclude Disposal Gain or Loss/Net Sales

X11 - Operating Expense Rate: Operating Expenses/Net Sales

X12 - Research and development expense rate: (Research and Development Expenses)/Net Sales

X13 - Cash flow rate: Cash Flow from Operating/Current Liabilities

X14 - Interest-bearing debt interest rate: Interest-bearing Debt/Equity

X15 - Tax rate (A): Effective Tax Rate

X16 - Net Value Per Share (B): Book Value Per Share(B)

X17 - Net Value Per Share (A): Book Value Per Share(A)

X18 - Net Value Per Share (C): Book Value Per Share(C)

X19 - Persistent EPS in the Last Four Seasons: EPS-Net Income

X20 - Cash Flow Per Share

X21 - Revenue Per Share (Yuan ¥): Sales Per Share

X22 - Operating Profit Per Share (Yuan ¥): Operating Income Per Share

X23 - Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share

X24 - Realized Sales Gross Profit Growth Rate

X25 - Operating Profit Growth Rate: Operating Income Growth

X26 - After-tax Net Profit Growth Rate: Net Income Growth

X27 - Regular Net Profit Growth Rate: Continuing Operating Income after Tax Growth

X28 - Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth

X29 - Total Asset Growth Rate: Total Asset Growth

X30 - Net Value Growth Rate: Total Equity Growth

X31 - Total Asset Return Growth Rate Ratio: Return on Total Asset Growth

X32 - Cash Reinvestment %: Cash Reinvestment Ratio

X33 - Current Ratio

X34 - Quick Ratio: Acid Test

X35 - Interest Expense Ratio: Interest Expenses/Total Revenue

X36 - Total debt/Total net worth: Total Liability/Equity Ratio

X37 - Debt ratio %: Liability/Total Assets

X38 - Net worth/Assets: Equity/Total Assets

X39 - Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets

X40 - Borrowing dependency: Cost of Interest-bearing Debt

X41 - Contingent liabilities/Net worth: Contingent Liability/Equity

X42 - Operating profit/Paid-in capital: Operating Income/Capital

X43 - Net profit before tax/Paid-in capital: Pretax Income/Capital

X44 - Inventory and accounts receivable/Net value: (Inventory+Accounts Receivables)/Equity

X45 - Total Asset Turnover

X46 - Accounts Receivable Turnover

X47 - Average Collection Days: Days Receivable Outstanding

X48 - Inventory Turnover Rate (times)

X49 - Fixed Assets Turnover Frequency

X50 - Net Worth Turnover Rate (times): Equity Turnover

X51 - Revenue per person: Sales Per Employee

X52 - Operating profit per person: Operation Income Per Employee

X53 - Allocation rate per person: Fixed Assets Per Employee

X54 - Working Capital to Total Assets

X55 - Quick Assets/Total Assets

X56 - Current Assets/Total Assets

X57 - Cash/Total Assets

X58 - Quick Assets/Current Liability

X59 - Cash/Current Liability

X60 - Current Liability to Assets

X61 - Operating Funds to Liability

X62 - Inventory/Working Capital

X63 - Inventory/Current Liability

X64 - Current Liabilities/Liability

X65 - Working Capital/Equity

X66 - Current Liabilities/Equity

X67 - Long-term Liability to Current Assets

X68 - Retained Earnings to Total Assets

X69 - Total income/Total expense

X70 - Total expense/Assets

X71 - Current Asset Turnover Rate: Current Assets to Sales

X72 - Quick Asset Turnover Rate: Quick Assets to Sales

X73 - Working capitcal Turnover Rate: Working Capital to Sale

X74 - Cash Turnover Rate: Cash to Sales

X75 - Cash Flow to Sales

X76 - Fixed Assets to Assets

X77 - Current Liability to Liability

X78 - Current Liability to Equity

X79 - Equity to Long-term Liability

X80 - Cash Flow to Total Assets

X81 - Cash Flow to Liability

X82 - CFO to Assets

X83 - Cash Flow to Equity

X84 - Current Liability to Current Assets

X85 - Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise

X86 - Net Income to Total Assets

X87 - Total assets to GNP price

X88 - No-credit Interval

X89 - Gross Profit to Sales

X90 - Net Income to Stockholder's Equity

X91 - Liability to Equity

X92 - Degree of Financial Leverage (DFL)

X93 - Interest Coverage Ratio (Interest expense to EBIT)

X94 - Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise

X95 - Equity to Liability

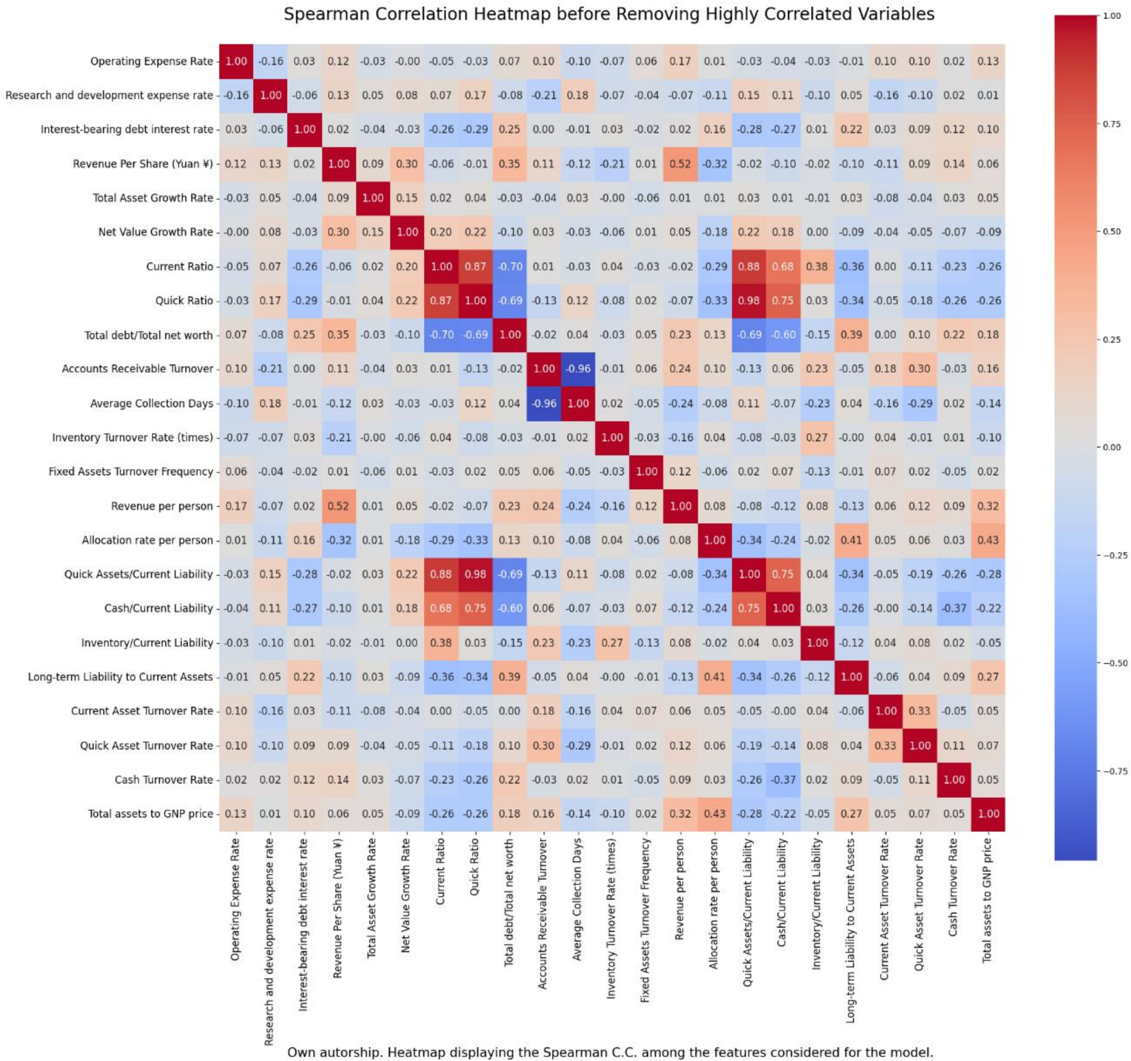
III. Data preprocessing and model deployment

a. XGBoost Classifier for bankruptcy prediction analysis.

Very little preprocessing was needed for analyzing the bankruptcy dataset. Most variables within the data set were already normalized and ranged from 0 to 1 for all values. All features were pre-engineered before the analysis and there were no NaN values detected. To build the best possible model required appropriate selection of important features as the leading analytical issue. My pipeline division included three major stages explained below.

I started by deleting features with a small variance since the normalized variables' threshold value was set at 0.1. By removing such features, the model enhances its performance as these components provide no valuable information and just introduce noise. Features which present zero variation across samples maintain a steady constant value thus they are unable to differentiate patterns or forecast new outcomes. The model retains predictive power after these irrelevant attributes are dropped from the analysis. When algorithms such as linear models or distance-based methods are used the presence of low-variance features generates mostly noise rather than signal while setting the model prone to overfitting. The model performs better at generalizing to new data after the removal of these unimportant features because it enhances its ability to detect key patterns. Model training along with inference runs faster using fewer features which benefits complex models during operation on large datasets. Anyways, if the reader want to see which variables were dropped, the second appendix contain a list of all of them, including those used in our model and those which were not considered (for any of the reasons explained here).

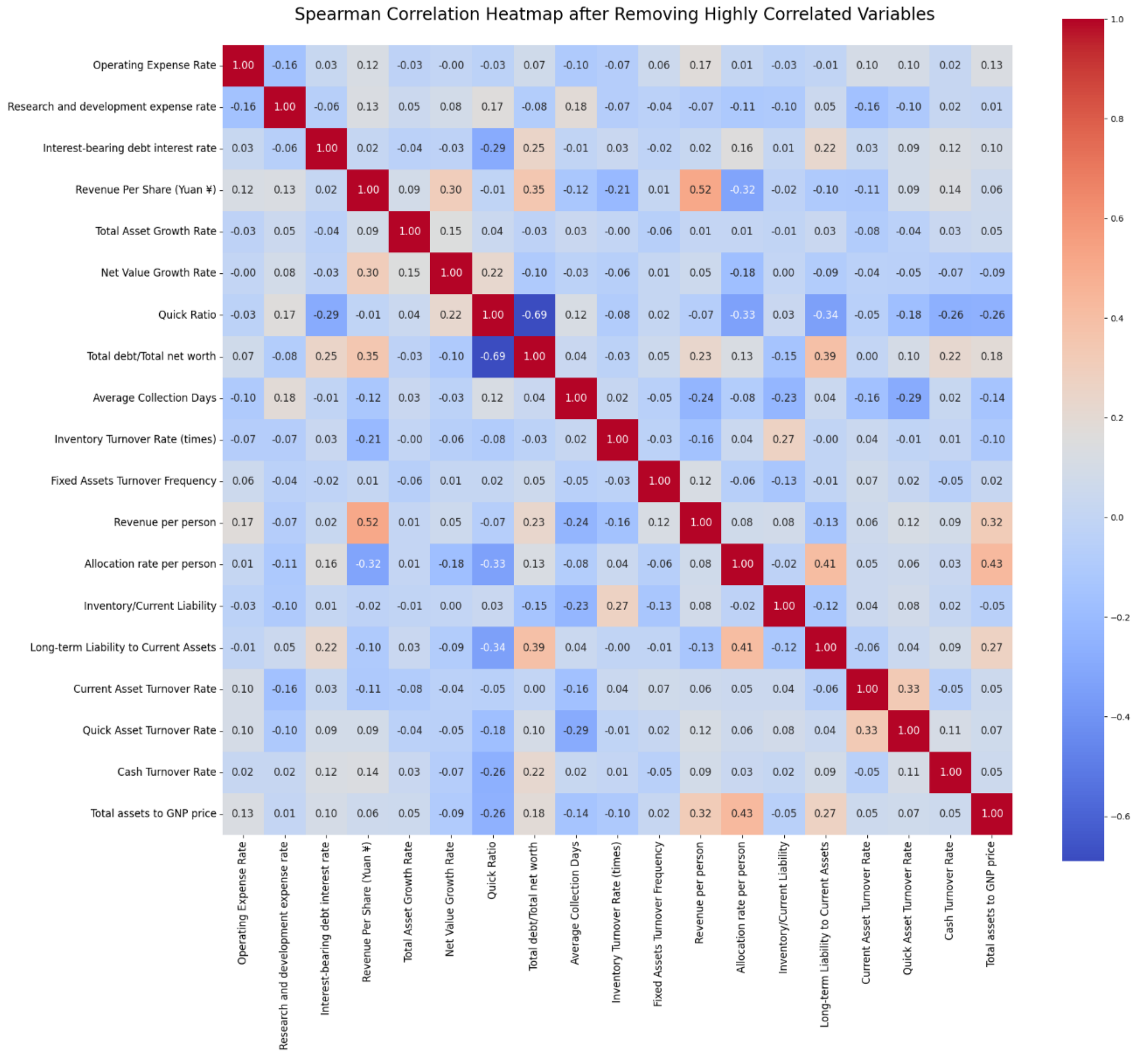
Secondly, I assessed potential correlations which existed between the features that remained. A second correlation heatmap was plotted for the purpose of avoiding negative model issues and bad performance caused by multicollinearity.



Own autorship. Heatmap displaying the Spearman C.C. among the features considered for the model. Figure deployed using Python (Matplotlib & Seaborn). The variables are normalized (0-1).

The presence of significant variable correlations throughout the data, limits efficient model development at each stage of the process due to their adverse effects on both interpretability and reliability as well as model performance. The model fails to distinguish distinct effects between correlated features because of which its coefficient estimates become unstable. The model becomes unreliable because of its inherent instability when small data changes cause performance to become unstable. The model loses effectiveness in generalizing beyond what its training data includes. Trouble emerges for determining true importance of features when similar predictor variables relate to each other because this connection distorts coefficient sizes and can reverse between positive and negative signs which makes important learning points difficult to understand.

After removing the highly correlated variables leaving in the model those who present the higher importance for the model according to the SHAP Importance and Total Gain importance thresholds (explained later), the final variables take this format in a heatmap. As we can observe, there is not any pair of variables which present a spearman correlation coefficient greater than 0.7 (our determined threshold).



	Operating Expense Rate	Research and development expense rate	Interest-bearing debt interest rate	Revenue Per Share (Yuan ¥)	Total Asset Growth Rate	Net Value Growth Rate	Quick Ratio	Total debt/Total net worth	Average Collection Days	Inventory Turnover Rate (times)	Fixed Assets Turnover Frequency	Revenue per person	Allocation rate per person	Inventory/Current Liability	Long-term Liability to Current Assets	Current Asset Turnover Rate	Quick Asset Turnover Rate	Cash Turnover Rate	Total assets to GNP price
Operating Expense Rate	1.00	-0.16	0.03	0.12	-0.03	-0.00	-0.03	0.07	-0.10	-0.07	0.06	0.17	0.01	-0.03	-0.01	0.10	0.10	0.02	0.13
Research and development expense rate	-0.16	1.00	-0.06	0.13	0.05	0.08	0.17	-0.08	0.18	-0.07	-0.04	-0.07	-0.11	-0.10	0.05	-0.16	-0.10	0.02	0.01
Interest-bearing debt interest rate	0.03	-0.06	1.00	0.02	-0.04	-0.03	-0.29	0.25	-0.01	0.03	-0.02	0.02	0.16	0.01	0.22	0.03	0.09	0.12	0.10
Revenue Per Share (Yuan ¥)	0.12	0.13	0.02	1.00	0.09	0.30	-0.01	0.35	-0.12	-0.21	0.01	0.52	-0.32	-0.02	-0.10	-0.11	0.09	0.14	0.06
Total Asset Growth Rate	-0.03	0.05	-0.04	0.09	1.00	0.15	0.04	-0.03	0.03	-0.00	-0.06	0.01	0.01	-0.01	0.03	-0.08	-0.04	0.03	0.05
Net Value Growth Rate	-0.00	0.08	-0.03	0.30	0.15	1.00	0.22	-0.10	-0.03	-0.06	0.01	0.05	-0.18	0.00	-0.09	-0.04	-0.05	-0.07	-0.09
Quick Ratio	-0.03	0.17	-0.29	-0.01	0.04	0.22	1.00	-0.69	0.12	-0.08	0.02	-0.07	-0.33	0.03	-0.34	-0.05	-0.18	-0.26	-0.26
Total debt/Total net worth	0.07	-0.08	0.25	0.35	-0.03	-0.10	-0.69	1.00	0.04	-0.03	0.05	0.23	0.13	-0.15	0.39	0.00	0.10	0.22	0.18
Average Collection Days	-0.10	0.18	-0.01	-0.12	0.03	-0.03	0.12	0.04	1.00	0.02	-0.05	-0.24	-0.08	-0.23	0.04	-0.16	-0.29	0.02	-0.14
Inventory Turnover Rate (times)	-0.07	-0.07	0.03	-0.21	-0.00	-0.06	-0.08	-0.03	0.02	1.00	-0.03	-0.16	0.04	0.27	-0.00	0.04	-0.01	0.01	-0.10
Fixed Assets Turnover Frequency	0.06	-0.04	-0.02	0.01	-0.06	0.01	0.02	0.05	-0.05	-0.03	1.00	0.12	-0.06	-0.13	-0.01	0.07	0.02	-0.05	0.02
Revenue per person	0.17	-0.07	0.02	0.52	0.01	0.05	-0.07	0.23	-0.24	-0.16	0.12	1.00	0.08	0.08	-0.13	0.06	0.12	0.09	0.32
Allocation rate per person	0.01	-0.11	0.16	-0.32	0.01	-0.18	-0.33	0.13	-0.08	0.04	-0.06	0.08	1.00	-0.02	0.41	0.05	0.06	0.03	0.43
Inventory/Current Liability	-0.03	-0.10	0.01	-0.02	-0.01	0.00	0.03	-0.15	-0.23	0.27	-0.13	0.08	-0.02	1.00	-0.12	0.04	0.08	0.02	-0.05
Long-term Liability to Current Assets	-0.01	0.05	0.22	-0.10	0.03	-0.09	-0.34	0.39	0.04	-0.00	-0.01	-0.13	0.41	-0.12	1.00	-0.06	0.04	0.09	0.27
Current Asset Turnover Rate	0.10	-0.16	0.03	-0.11	-0.08	-0.04	-0.05	0.00	-0.16	0.04	0.07	0.06	0.05	0.04	-0.06	1.00	0.33	-0.05	0.05
Quick Asset Turnover Rate	0.10	-0.10	0.09	0.09	-0.04	-0.05	-0.18	0.10	-0.29	-0.01	0.02	0.12	0.06	0.08	0.04	0.33	1.00	0.11	0.07
Cash Turnover Rate	0.02	0.02	0.12	0.14	0.03	-0.07	-0.26	0.22	0.02	0.01	-0.05	0.09	0.03	0.02	0.09	-0.05	0.11	1.00	0.05
Total assets to GNP price	0.13	0.01	0.10	0.06	0.05	-0.09	-0.26	0.18	-0.14	-0.10	0.02	0.32	0.43	-0.05	0.27	0.05	0.07	0.05	1.00

Own autorship. Heatmap displaying the Spearman C.C. among the features considered for the model. Figure deployed using Python (Matplotlib & Seaborn). The variables are normalized (0-1).

These variables function within multiple financial categories because of which they represent essential metrics for analyzing complete financial performance of companies. The described fields of financial performance classify into specific categories.

Operating efficiency tracking with cost management enables companies to disclose vital data regarding their profitability and sustainability performance. Operating Expense Rates help firms check their cost-to-revenue ratio because elevated expenses create direct reduction in their profit potentials. The Research and Development Expense Rate shows that a company firmly supports innovation while preserving its market position. The deliberate R&D investment of companies brings them better prospects for future growth. All businesses irrespective of size can use Revenue Per Share to evaluate their revenue levels since this performance measure expresses earnings in standardized units regardless of stock ownership.

The evaluation of leverage through specific methods provides companies with debt sustainability metrics needed to understand their financial risks and capital structure design. Better creditworthiness indicates a low Interest-Bearing Debt Interest Rate because it indicates reduced borrowing expenses. A company's solvency position can be determined by analyzing its Total Debt to Total Net Worth ratio because this metric shows both dependence on debt and complete company value. Financial management contains both single-ratio analysis and pattern-based assessments of multiple metrics. Organizations should analyze their long-term debt relationship to short-term assets through the Long-term Liability to Current Assets ratio to measure financial stability levels.

Businesses need vital temporary financial health and liquidity indicators for determining their capacity to fulfill current obligations. Businesses should use the Quick Ratio because it shows comprehensive information about immediate solvency without inventory. Manufacturing and retail sector companies must use the Inventory to Current Liability ratio to determine if their inventory management aligns properly with their short-term debt requirements. Average Collection Days enables companies to evaluate receivables efficiency by monitoring extended payment durations which causes cash flow issues that signal funding stability risks.

These indicators show the productive use of organizational assets to create revenue streams. A company achieves strong sales results from high Inventory Turnover Rate while low rates signify either overstocked or obsolete inventory. Through the

Fixed Assets Turnover Frequency assessment organizations can determine their ability to produce revenue from their long-term property and equipment investments. Alongside the Current Asset Turnover Rate and Quick Asset Turnover Rate a company can understand its operational agility by observing how liquid assets contribute to sales generation.

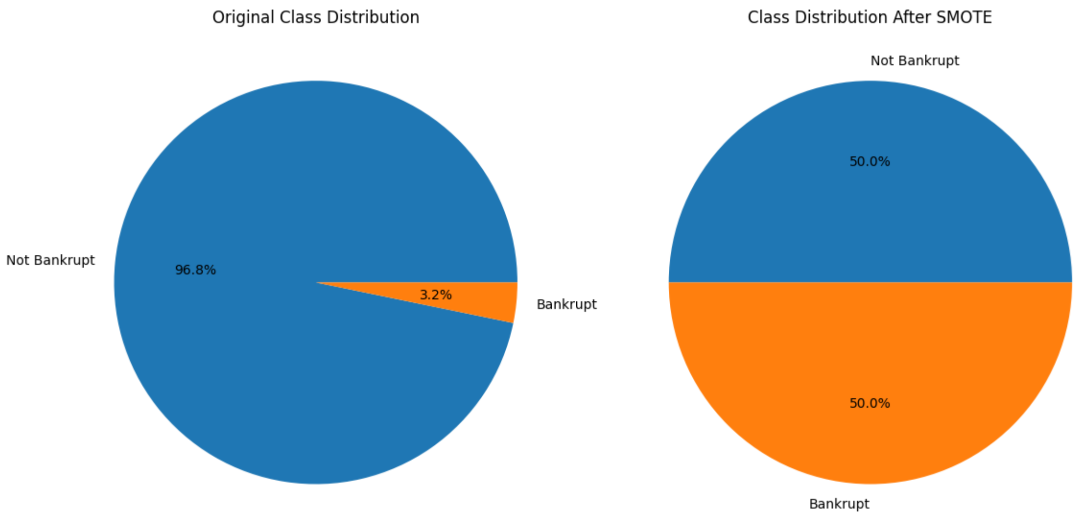
Growth and productivity metrics: these provide a forward-looking perspective on a company's expansion and operational effectiveness. Firms can evaluate resource expansion and shareholder value growth through measures of Total Asset Growth Rate and Net Value Growth Rate. The workforce efficiency measures Revenue per Person together with Allocation Rate per Person are key metrics in industries that heavily rely on human capital expenses for cost management.

The macro-financial linkages metrics enable companies to understand their performance relations with economic conditions. Total Assets to GNP Price ratio brings company-level information to macroeconomic factors while providing useful insights into potential economic risks.

After checking for no mathematical and technical-related issues on our model, and checking that all the variables to be included actually make sense under the financial theoretical framework, we can conclude our feature selection process for the first model.

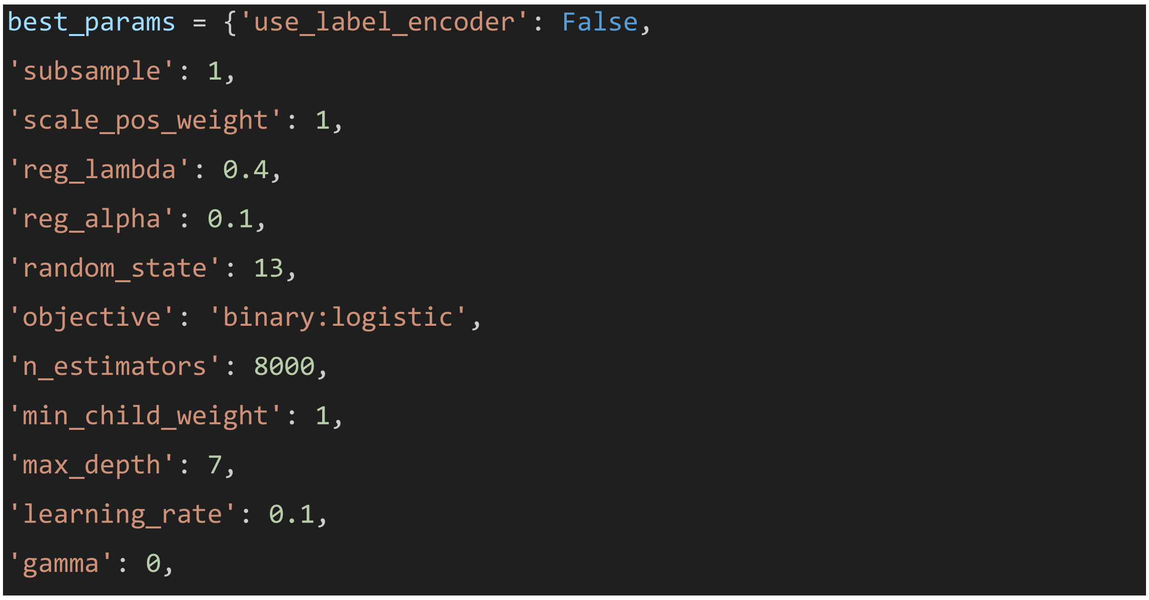
Before model deployment I examined the target variable "Bankrupt?" which displays an unbalanced nature. A dataset presents class imbalance when it displays heavy skewed distribution between its classes resulting in significant differences between class frequencies. A model trained on imbalanced data develops a bias toward major class examples because optimization for total accuracy results in limited learning of minority class patterns.

The best solution to tackle this problem involves utilizing SMOTE (Synthetic Minority Over-sampling Technique) method. The main advantage of SMOTE over traditional oversampling methods becomes clear since it differs from minor-class sample duplication by creating artificial examples for better dataset equilibrium. The methodology starts by picking one minority-class example and determines the k-nearest neighbors located within the feature space. The system determines new synthetic points that exist between original samples and their neighboring examples. This process led us to achieve this newly developed distribution.



Own authorship. Left: Pie plot displaying class imbalance in the target variable.
Right: Balanced classes after applying SMOTE (Oversampling Technique).
Images created with Python (matplotlib).

After this adjustment, the model was deployed. Before talking about the results, we have to talk about the hyperparameter tunning process carried out. Particularly, the parameters used to configure the model, are the following:



```
best_params = {'use_label_encoder': False,
'subsample': 1,
'scale_pos_weight': 1,
'reg_lambda': 0.4,
'reg_alpha': 0.1,
'random_state': 13,
'objective': 'binary:logistic',
'n_estimators': 8000,
'min_child_weight': 1,
'max_depth': 7,
'learning_rate': 0.1,
'gamma': 0,
'eval_metric': 'aucpr',
'colsample_bytree': 1}
```

The process achieved its targets mainly through Randomized Search for initial threshold discovery while manual testing sought appropriate final values for increased precision and reduced overfitting risks. The absence of grid search capabilities due to limited computing resources led me to select the mentioned solution since I deemed it most appropriate across available options.

All these technical settings determine how the model processes data along with complexity-management mechanisms and final performance levels between training data along with unknown inputs. Different parameters in the model establish relationships which modify the model's prediction quality and capability to protect against overfitting.

The objective function refers to binary logistic as the model generates probability estimates instead of producing direct class outputs. The selection of this model works well for systems that depend on confidence evaluation and trained decision threshold establishment. Model probability outputs prove highly beneficial for different decision-making contexts that need prediction ranking or assessment of performance over multiple confidence levels. My choice for evaluation was to use the precision-recall curve area under the curve as the main metric. During optimization this approach ensures the model maintains steady controls of precision and recall at every prediction threshold point. This metric provides essential assessment for imbalanced classification situations because it allows identification of positive cases accurately alongside reasonable reduction of false positive errors.

The boosting rounds specify the number of sequential trees that will be built in a sequence to fix previous mistakes. The number of boosting rounds enables the model to refine its predictions better with additional adjustments, but the net advantage relies on the learning rate and early stopping parameters. Each new decision tree contributes less to the prediction during training through the learning rate adjustment which needs multiple trees for optimal generalization results.

The parameters shape how tree complexity functions in this model. Maximum depth serves as a limit for tree depth which determines the complexity of patterns the model detects. The minimum child weight parameter defines the necessary conditions for partition creation to stop the model from splitting instances scarcely. The splitting operation in the Gamma method needs a minimum amount of loss improvement before proceeding with a split.

The L1 as well as L2 regularization terms serve to implement this technique. The L1 regularization element in the model pushes some weights to zero to achieve sparsity while simultaneously selecting important features. L2 regularization distributes the feature weights relatively evenly to stop any one feature from dominating others. The combination of L1 and L2 regularization terms works jointly to avoid overfitting without affecting predictive capacity.

The subsampling parameters enable systematic random selection of both training samples together with features when constructing trees. Tomography through this technique generates varied trees which leads to better generalization. Model behavior when dealing with imbalanced datasets is modified using the positive class weight scale parameter to control example importance levels.

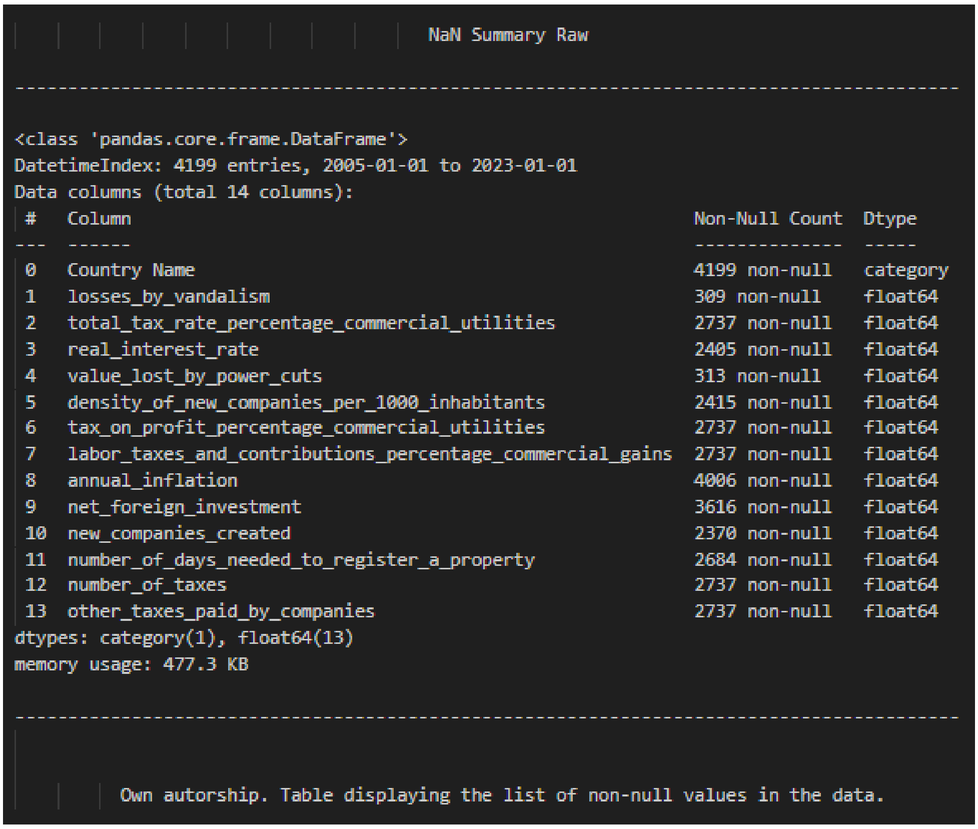
b. XGBoost Regressor for firm creation/destruction prediction analysis.

More preprocessing, compared to the previous model, was needed for the analysis and prediction of firm creation/destruction. Reduced the number of features compared to bankruptcy data simplified the selection process because we now have the capability to test various combinations manually in order to achieve optimal results. In addition to default variables, we engineered several new variables as part of feature engineering including continent-based data and crisis information.

As mentioned at the beginning, the World Bank Data Center provides trustworthy data, but we need to evaluate the absence of complete values and unusual data patterns which affect the data quality. Some data-handling procedures enable us to improve the current quality standards.

The data contains different amounts of null values among their variables as shown by this table. The technique used for filling these null values will consist of interpolation, which provides an appropriate solution because we work with time series data. The methodology helps determine values in sequences made of known measurements. Time series analysis benefits strongly from interpolation because this method allows us to correctly fill missing (NaN) values which conform to the underlying temporal framework of the data. The time-aware interpolation method conquers record incompleteness without linear simplification because it protects generated values from shifting outside their correct time slots based on the real observation time spans of records.

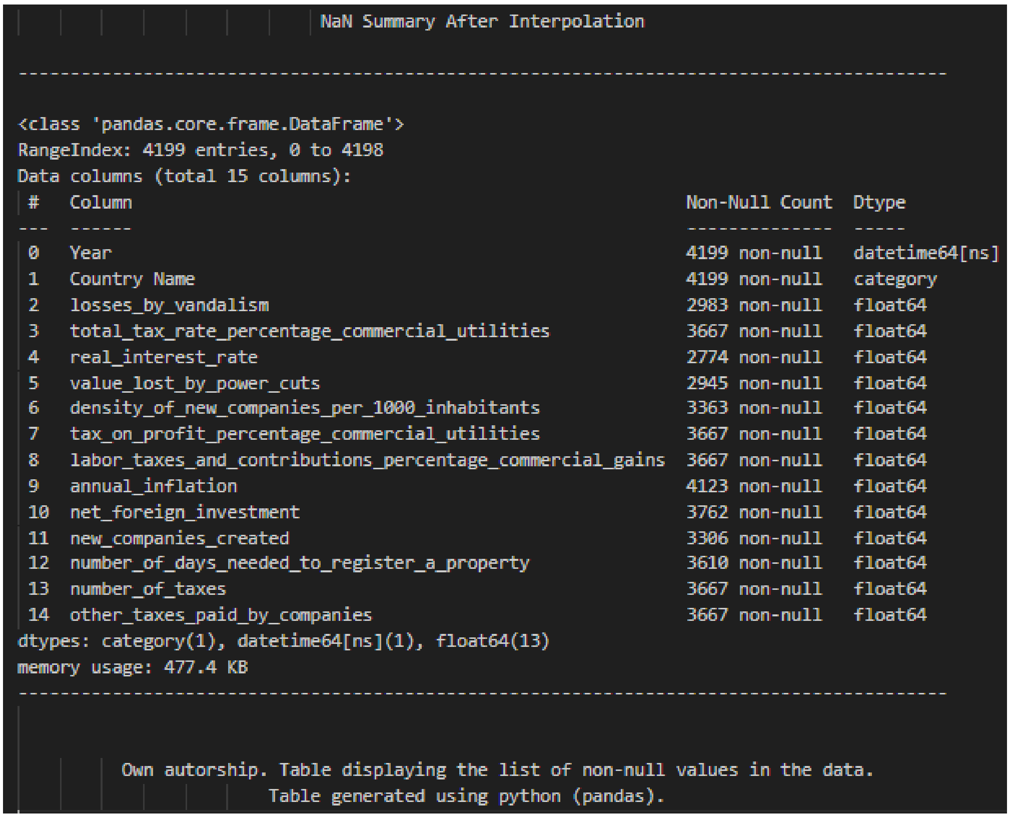
I applied pandas’ interpolate(method='time') function as part of my solution to interpolate data by referencing the DataFrame time index. This method fits well with datasets that have irregular distribution of observations since it adjusts the interpolation weights according to temporal distances between measurement points. The application of bidirectional interpolation through 'limit_direction='both'' enables the propagation of known values into forward and backward direction across each group. Such a method manages gaps whether they occur at the initial or terminal parts of time series data before unidirectional methods would keep NaN values in place. Group-wise interpolation is used for each country separately because applying inference from unrelated regions would create contamination of the dataset. Sorting the data by time index followed by grouping by 'Country Name' ensures that the interpolation algorithm maintains appropriate chronological order as well as regional separation of the data points. Time series integrity remains intact from our imputation approach because the strategy supplies localized replacements that reduce potential bias introduced through other non-temporal or global replacement methods. The technique appealed to me because it establishes a balanced relationship between running time optimization and maintaining the core data relationships. Time-aware interpolation shows robustness as a preprocessing step for machine learning applications because it fits various local trends in each country's time series data as an alternative to basic methods like zero-filling or mean imputation.



```
                                        NaN Summary Raw

--------------------------------------------------------------------------------------------

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 4199 entries, 2005-01-01 to 2023-01-01
Data columns (total 14 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   Country Name                                          4199 non-null   category
 1   losses_by_vandalism                                   309 non-null    float64
 2   total_tax_rate_percentage_commercial_utilities        2737 non-null   float64
 3   real_interest_rate                                    2405 non-null   float64
 4   value_lost_by_power_cuts                              313 non-null    float64
 5   density_of_new_companies_per_1000_inhabitants         2415 non-null   float64
 6   tax_on_profit_percentage_commercial_utilities         2737 non-null   float64
 7   labor_taxes_and_contributions_percentage_commercial_gains 2737 non-null   float64
 8   annual_inflation                                      4006 non-null   float64
 9   net_foreign_investment                                3616 non-null   float64
 10  new_companies_created                                 2370 non-null   float64
 11  number_of_days_needed_to_register_a_property          2684 non-null   float64
 12  number_of_taxes                                       2737 non-null   float64
 13  other_taxes_paid_by_companies                         2737 non-null   float64
dtypes: category(1), float64(13)
memory usage: 477.3 KB

--------------------------------------------------------------------------------------------
```

Own autorship. Table displaying the list of non-null values in the data.

The procedure results in this output. We have gained considerable benefits because the estimated value of variables like losses by vandalism and value lost by power cuts now provides useful information for our analysis even though the data quality formerly presented significant limitations. Will the NaN values remain unaddressed in the dataset? Keeping those observations in our dataset is justified because XGBoost shows natural resistance to outliers that eliminates active outlier removal needs. Extreme data points impact less on XGBoost-based tree methods through its data partitioning approach when information gets spread across defined thresholds. The collective use of multiple weak learnings through ensemble prediction minimizes the overall effects that outliers have on individual tree splits. The method of holding all data points maintains their completeness including outliers because this prevents discarding any essential information. The removal of outliers may result in the removal of entire records when some countries have limited data observations thus reducing both the general coverage and data diversity. The inclusion of outliers in the dataset becomes important because some variables already suffer from missing value issues which resulted in incompleteness within certain countries.

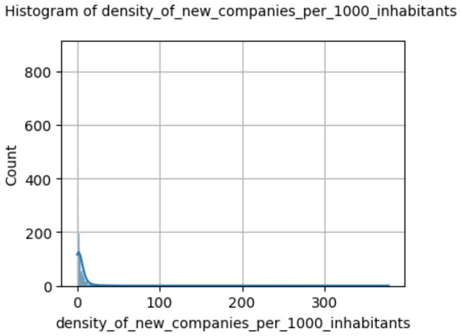


NaN Summary After Interpolation

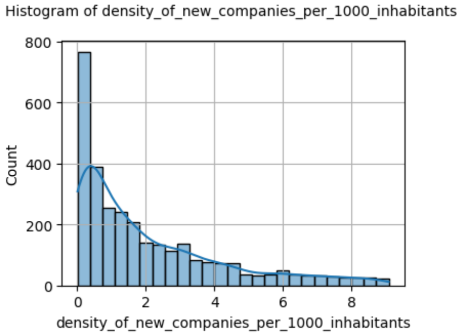
```
------------------------------------------------------------------------------------------

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4199 entries, 0 to 4198
Data columns (total 15 columns):
 #   Column                                            Non-Null Count  Dtype
---  ------                                            --------------  -----
 0   Year                                              4199 non-null   datetime64[ns]
 1   Country Name                                      4199 non-null   category
 2   losses_by_vandalism                               2983 non-null   float64
 3   total_tax_rate_percentage_commercial_utilities    3667 non-null   float64
 4   real_interest_rate                                2774 non-null   float64
 5   value_lost_by_power_cuts                          2945 non-null   float64
 6   density_of_new_companies_per_1000_inhabitants     3363 non-null   float64
 7   tax_on_profit_percentage_commercial_utilities     3667 non-null   float64
 8   labor_taxes_and_contributions_percentage_commercial_gains  3667 non-null   float64
 9   annual_inflation                                  4123 non-null   float64
 10  net_foreign_investment                            3762 non-null   float64
 11  new_companies_created                             3306 non-null   float64
 12  number_of_days_needed_to_register_a_property      3610 non-null   float64
 13  number_of_taxes                                   3667 non-null   float64
 14  other_taxes_paid_by_companies                     3667 non-null   float64
dtypes: category(1), datetime64[ns](1), float64(13)
memory usage: 477.4 KB
------------------------------------------------------------------------------------------
```

Own autorship. Table displaying the list of non-null values in the data.
Table generated using python (pandas).

To address critical data preparation elements I selected a selective method of dealing with outliers which targets removing only target variable anomalies while keeping their counterparts in the feature dimension. A dual strategy based on practical and theoretical factors led to this decision because it delivered superior model performance while keeping data retention to a minimum. The presence of outliers in target values produces training distortions that specifically affect regression models and unbalanced classification problems because these extreme values disrupt model bias performance. The model might select to fit anomalous response values that result in poor overall generalization because of their extreme magnitude. The removal of outliers in the target variable results in improved model predictive accuracy together with increased stability because the model receives training from a more balanced distribution. The main purpose of keeping feature variable outliers intact consisted of two important points. The inherent robustness of XGBoost bare ensemble methods allows this model to properly understand nonlinear patterns and handle extremely high or low variable values. The split-based decision mechanism of this model automatically reduces global effects of outliers by routing them into separate branches that prevent overall contamination. The features contain numerous cases which demonstrate rare-but-relevant anomalous patterns so they should not be automatically eliminated as outliers. The removal of such points might diminish the model's sophistication by eliminating important cases which display data variability. The model avoids unrealistic limitations in the real world when it maintains these points since it becomes more effective at actual field scenarios. The changes to target variable distributions because of outlier deletion become visible through the following illustrations.



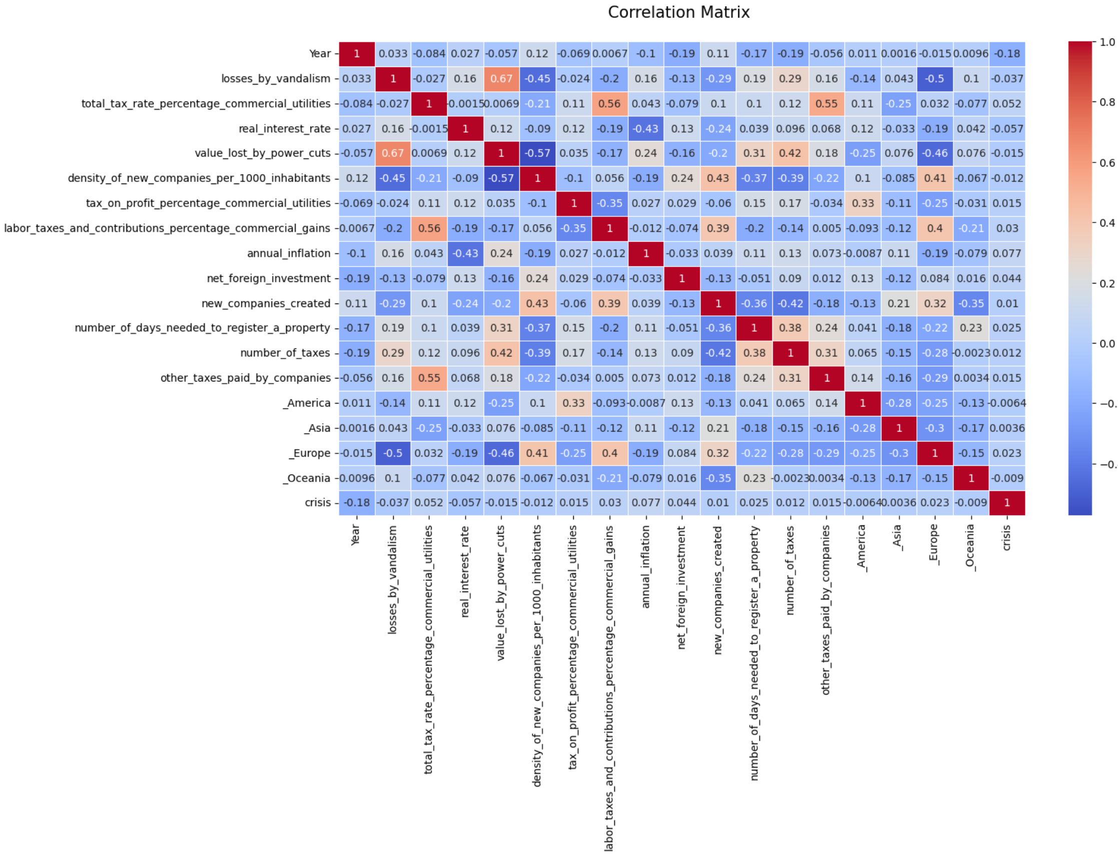
Own authorship. Histogram depicting the distribution of the target variable before dropping outliers with the IQR method. Image generated with python (matplotlib & seaborn).



Own authorship. Histogram depicting the distribution of the target variable after dropping outliers with the IQR method. Image generated with python (matplotlib & seaborn).

It is worth to mention that the I tested the model without removing outliers and still, it gave good results, quite similar to the ones obtained that will be shown later in the paper.

Finally, in order to conclude the preprocessing and feature selection process, we have to check once again for the correlations among the features that will be used to predict our target variable. As we can observe in the heatmap displayed below, in this case it was not necessary to drop any of our variables since it does not seem to be any correlation coefficient above our threshold (0.7).



	Year	losses_by_vandalism	total_tax_rate_percentage_commercial_utilities	real_interest_rate	value_lost_by_power_cuts	density_of_new_companies_per_1000_inhabitants	tax_on_profit_percentage_commercial_utilities	labor_taxes_and_contributions_percentage_commercial_gains	annual_inflation	net_foreign_investment	new_companies_created	number_of_days_needed_to_register_a_property	number_of_taxes	other_taxes_paid_by_companies	_America	_Asia	_Europe	_Oceania	crisis
Year	1	0.033	-0.084	0.027	-0.057	0.12	-0.069	0.0067	-0.1	-0.19	0.11	-0.17	-0.19	-0.056	0.011	0.0016	-0.015	0.0096	-0.18
losses_by_vandalism	0.033	1	-0.027	0.16	0.67	-0.45	-0.024	-0.2	0.16	-0.13	-0.29	0.19	0.29	0.16	-0.14	0.043	-0.5	0.1	-0.037
total_tax_rate_percentage_commercial_utilities	-0.084	-0.027	1	-0.0015	0.0069	-0.21	0.11	0.56	0.043	-0.079	0.1	0.1	0.12	0.55	0.11	-0.25	0.032	-0.077	0.052
real_interest_rate	0.027	0.16	-0.0015	1	0.12	-0.09	0.12	-0.19	-0.43	0.13	-0.24	0.039	0.096	0.068	0.12	-0.033	-0.19	0.042	-0.057
value_lost_by_power_cuts	-0.057	0.67	0.0069	0.12	1	-0.57	0.035	-0.17	0.24	-0.16	-0.2	0.31	0.42	0.18	-0.25	0.076	-0.46	0.076	-0.015
density_of_new_companies_per_1000_inhabitants	0.12	-0.45	-0.21	-0.09	-0.57	1	-0.1	0.056	-0.19	0.24	0.43	-0.37	-0.39	-0.22	0.1	-0.085	0.41	-0.067	-0.012
tax_on_profit_percentage_commercial_utilities	-0.069	-0.024	0.11	0.12	0.035	-0.1	1	-0.35	0.027	0.029	-0.06	0.15	0.17	-0.034	0.33	-0.11	-0.25	-0.031	0.015
labor_taxes_and_contributions_percentage_commercial_gains	0.0067	-0.2	0.56	-0.19	-0.17	0.056	-0.35	1	-0.012	-0.074	0.39	-0.2	-0.14	0.005	-0.093	-0.12	0.4	-0.21	0.03
annual_inflation	-0.1	0.16	0.043	-0.43	0.24	-0.19	0.027	-0.012	1	-0.033	0.039	0.11	0.13	0.073	-0.0087	0.11	-0.19	-0.079	0.077
net_foreign_investment	-0.19	-0.13	-0.079	0.13	-0.16	0.24	0.029	-0.074	-0.033	1	-0.13	-0.051	0.09	0.012	0.13	-0.12	0.084	0.016	0.044
new_companies_created	0.11	-0.29	0.1	-0.24	-0.2	0.43	-0.06	0.39	0.039	-0.13	1	-0.36	-0.42	-0.18	-0.13	0.21	0.32	-0.35	0.01
number_of_days_needed_to_register_a_property	-0.17	0.19	0.1	0.039	0.31	-0.37	0.15	-0.2	0.11	-0.051	-0.36	1	0.38	0.24	0.041	-0.18	-0.22	0.23	0.025
number_of_taxes	-0.19	0.29	0.12	0.096	0.42	-0.39	0.17	-0.14	0.13	0.09	-0.42	0.38	1	0.31	0.065	-0.15	-0.28	-0.0023	0.012
other_taxes_paid_by_companies	-0.056	0.16	0.55	0.068	0.18	-0.22	-0.034	0.005	0.073	0.012	-0.18	0.24	0.31	1	0.14	-0.16	-0.29	0.0034	0.015
_America	0.011	-0.14	0.11	0.12	-0.25	0.1	0.33	-0.093	-0.0087	0.13	-0.13	0.041	0.065	0.14	1	-0.28	-0.25	-0.13	-0.0064
_Asia	0.0016	0.043	-0.25	-0.033	0.076	-0.085	-0.11	-0.12	0.11	-0.12	0.21	-0.18	-0.15	-0.16	-0.28	1	-0.3	-0.17	0.0036
_Europe	-0.015	-0.5	0.032	-0.19	-0.46	0.41	-0.25	0.4	-0.19	0.084	0.32	-0.22	-0.28	-0.29	-0.25	-0.3	1	-0.15	0.023
_Oceania	0.0096	0.1	-0.077	0.042	0.076	-0.067	-0.031	-0.21	-0.079	0.016	-0.35	0.23	-0.0023	0.0034	-0.13	-0.17	-0.15	1	-0.009
crisis	-0.18	-0.037	0.052	-0.057	-0.015	-0.012	0.015	0.03	0.077	0.044	0.01	0.025	0.012	0.015	-0.0064	0.0036	0.023	-0.009	1

Own autorship. Heatmap displaying the Spearman C.C. among the features considered for the model.
Figure deployed using Python (Matplotlib & Seaborn).

8. Bibliography

- World Bank. (2025a). *Densidad de nuevas empresas cada 1000...* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025b). *Impuesto sobre utilidades porcentaje util...* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025c). *Impuestos laborales y contribuciones por...* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025d). *Inflacion Anual* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025e). *Inversion Extrangera Neta - World Bank* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025f). *Nuevas Empresas Creadas - Banco Mund...* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025g). *Numero de días necesarios para registrar ...* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025h). *Numero de impuestos* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025i). *Otros impuestos que pagan las empresas* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025j). *Perdidas por Robo, Vandalismo, Asalto e ...* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025k). *Tasa tributaria total porcentage de utilida...* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025l). *Tipos de Interes Real* [Data set]. World Bank Database. https://databank.worldbank.org
- World Bank. (2025m). *Valor perdido por cortes electrices - Worl...* [Data set]. World Bank Database. https://databank.worldbank.org
- **Segovia Vargas, M. J., & Camacho Miñano, M. del M.** (2018). Analysis of corporate viability in the pre-bankruptcy proceedings. *Contaduría y Administración, 63*(1), 1–17.
- **Adamko, P., & Chutka, J.** (2020). Company bankruptcy and its prediction in conditions of globalization. *SHS Web of Conferences, 74*, 05002.
- **Kadarningsih, A., Oktavia, V., Falah, T. R., & Sari, Y. S.** (2021). Profitability as determining factor to anticipate company bankruptcy. *Estudios de Economía Aplicada, 39*(12).
- Fede Soriano. (2021). *Company Bankruptcy Prediction* [Data set]. Kaggle. https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction

- **Segovia Vargas, M. J., & Camacho Miñano, M. del M.** (2018). *Analysis of corporate viability in the pre-bankruptcy proceedings. Contaduría y Administración*, 63(1), 1–*17*. https://doi.org/10.22201/fca.24488410e.2018.1022
- **Hundt, C., & Sternberg, R.** (2016). *Explaining new firm creation in Europe from a spatial and time perspective: A multilevel analysis based upon data of individuals, regions and countries*. Papers in Regional Science, 95(1), 1–36. https://doi.org/10.1111/pirs.12133