Borsa Istanbul Review 24 (2024) 47–60

Contents lists available at ScienceDirect

Borsa Istanbul Review

journal homepage: www.elsevier.com/journals/borsa-istanbul-review/2214-8450

The good, the better and the challenging: Insights into predicting high-growth firms using machine learning

Sermet Pekin *, Aykut Şengül

Central Bank of the Republic of Türkiye, Research and Monetary Policy Department, Türkiye

ARTICLE INFO

JEL classification:
C40
C55
C60
C81
L25

Keywords:
High-growth firms
Machine learning
Prediction
Firm dynamics

ABSTRACT

This study aims to classify high-growth firms using several machine learning algorithms, including K-Nearest Neighbors, Logistic Regression with L1 (Lasso) and L2 (Ridge) Regularization, XGBoost, Gradient Descent, Naive Bayes and Random Forest. Leveraging a dataset composed of financial metrics and firm characteristics between 2009 and 2022 with 1,318,799 unique firms (averaging 554,178 annually), we evaluate the performance of each model using metrics such as MCC, ROC AUC, accuracy, precision, recall and F1-score. In our study, ROC AUC values ranged from 0.53 to 0.87 for employee-high growth and from 0.53 to 0.91 for turnover-high growth, depending on the method used. Our findings indicate that XGBoost achieves the highest performance, followed by Random Forest and Logistic Regression, demonstrating their effectiveness in distinguishing between high-growth and non-high-growth firms. Conversely, KNN and Naive Bayes yield lower accuracy. Furthermore, our findings reveal that growth opportunity emerges as the most significant factor in our study. This research contributes valuable insights to financial analysts and investors in identifying high-growth firms and underscores the potential of machine learning in economic prediction.

1. Introduction

Identifying high-growth firms (HGFs) is crucial for investors, analysts and policymakers, as these firms significantly contribute to job creation (Storey, 1994), innovation (Mason & Brown, 2012) and economic growth (Gupta et al., 2013). Amid the complexities of financial data, traditional regression models, reliant on financial ratios, often fail to capture the non-linear relationships inherent in this domain. Recent advancements in machine learning have introduced powerful algorithms capable of leveraging extensive datasets to enhance growth predictions.

HGFs have continued to attract significant attention in the research community. A recent Google Scholar search for *"high-growth firms"* shows over 25,600 results today. Similarly, Henrekson and Johansson's (2010) literature review has gained considerable attention. According to Google Scholar, it has gained 1320 additional citations, compared to the 251 citations reported by Coad et al. (2014).

Despite some progress in financial forecasting using machine learning, many studies have yielded unsatisfactory results. For instance, previous analyses often report low R^2 values and AUC scores, indicating limited predictive power (e.g., Srhoj, 2022). By incorporating firm-specific and time-varying variables alongside financial statement data, our study aims to bolster the accuracy of predictions using various machine learning techniques.

Models predicting firm growth by taking firm-specific characteristics into account focus on balance sheet data and argue that off-balance sheet factors should also be incorporated to increase the explanatory power of the forecast. Among the studies using financial statements, Srhoj (2022) predicts growth with an R^2 of 0.05 and an accuracy of 0.59 (AUC 0.63); Weinblat (2018) predicts growth with an AUC of 0.58 and an accuracy of 0.63; Coad and Srhoj (2020) predict with an R^2 of 0.10 and an accuracy of 0.77; VanWitteloostuijn and Kolkman (2019) predict with a minimum R^2 of 0.16 values. Because firm growth is a time-varying process (Penrose, 2009), besides balance sheet data, adding firm-specific and time-varying variables will increase the power of the estimation. In this context, the predictive power of the model in our study has been enhanced by adding variables such as firm productivity, high R&D investments, firm types (incorporated, limited, etc.), firm size, firm age and other time-varying variables. In addition, a validation test was conducted as an advantage of having a large dataset, which was missing in previous studies. In previous studies, limited tests were conducted to increase the predictive power and growth ratios of firms

* Corresponding author.
E-mail addresses: sermet.pekin@tcmb.gov.tr (S. Pekin), aykut.sengul@tcmb.gov.tr (A. Şengül).

https://doi.org/10.1016/j.bir.2024.12.001
Received 3 September 2024; Received in revised form 17 December 2024; Accepted 17 December 2024
Available online 18 December 2024

were also included; in our study, no direct variables related to employee, turnover or asset growth were included. The superiority of machine learning prediction over traditional models, the various robust tests and the different estimation methods used highlight the robustness of our results.

The remainder of the paper is organized as follows: Section 2 analyzes the literature primarily focusing on ML methodology and Section 3 presents some insight on input and output variables and methodology. Section 4 presents the results of the different prediction methods and robustness tests. Section 5 discusses our findings in the literature and the superiority of our models. Lastly, Section 6 concludes the paper.

2. Literature review

Despite numerous studies aimed at understanding the dynamics of firm growth, the results regarding firm growth in the literature have often remained contradictory and uncertain. The first studies on firm growth were the classical Gibrat law (Gibrat, 1931), which states that firm growth is random. Subsequent studies on firm growth can be grouped into three categories: first studies argue that internal firm-specific heterogeneous changes (Storey, 1994; Acs et al., 2008) cause firm growth. Second studies consider external factors causing firm growth (Capon et al., 1990; Davidsson et al., 2010). Third studies state that macroeconomic cycles cause firm growth (Geroski et al., 2010).

Investigating the causes of firm growth has led to the research and prediction of high-growth firms. The heterogeneous nature of firms (Delmar et al., 2003) makes it difficult to draw precise inferences about predicting high growth. Furthermore, Audretsch and Dohse (2007) emphasize that one of the greatest hindrances to analyzing the relationship between firm growth and its drivers has been the lack of access to longitudinal datasets. In addition to studies that argue growth is purely random and cannot be predicted precisely (Coad et al., 2013) and that luck is a significant factor for high-growth firms (Storey, 2011), some studies argue that HGFs can be predicted.

Among these studies, linear regression (Hölzl, 2014) and quantile regression (Sampagnaro & Lubrano Lavadera, 2013) are used as conventional methods to identify the characteristics of HGFs. Probit regression (Daunfeldt & Halvarsson, 2014) and logit models (Segarra & Teruel, 2014) are also employed to predict whether firms will be HGFs. Traditional methods like Lasso (Chae, 2024; Coad & Srhoj, 2020) and Logistic Regression (Guzman & Stern, 2020) have limitations in detecting nonlinear relationships. Consequently, the low success of traditional methods in predicting firm growth (Coad, 2009) and their lack of explanatory power (Marsili, 2001) have motivated scholars to explore big data and machine learning methods to identify HGFs.

In our study, we account for firm-specific heterogeneous factors and institutional dynamics to model the characteristics that increase the probability of high growth. Türkiye, as a transition economy with high national income growth and a high unemployment rate, presents a critical context for understanding firm dynamics. Our large micro dataset combined with machine learning methods seeks to increase the reliability of predictions compared to previous studies, emphasizing the importance of various factors in firm growth.

We differ from past methodologies in several ways. First, we examine the definitions of employee and turnover growth, along with asset growth, in accordance with OECD standards. Second, we apply three growth indicators using seven different methods and employ stratified sector-specific (manufacturing) and narrowed models to mitigate the risk of spurious results and enhance robustness. Third, we added variables that contain the firm network relationship by using the declaration of purchase and declaration of sale dataset for HGF estimation for the first time in the literature. Recent research has begun to explore machine learning applications in financial prediction, demonstrating the potential of algorithms such as Random Forest, Gradient Boosting and Logistic Regression to outperform conventional methods. However, a comprehensive comparative analysis evaluating these models in predicting HGFs is still lacking. This study addresses this gap by systematically comparing multiple machine learning algorithms on a common dataset, yielding valuable insights into their relative performance.

3. Methodology

3.1. Target variables

Academic research has analyzed firm growth as a change in specific parameters, such as quantitative differences (Davidsson et al., 2010). A well-accepted measure of dynamism and competitiveness for a firm is turnover growth. Although turnover is sensitive to inflation and exchange rates, it is commonly used to measure firm growth. In addition to turnover, employment is often used due to its ease of access, straightforward measurement and relevance to policymakers. The Eurostat-OECD Guide to Business Demographics Statistics (Eurostat-OECD, 2008) also recommends considering employment and turnover in analyzing fast-growing enterprises. Furthermore, examining asset growth will also help confirm the consistency of the results. Therefore, all three variables are analyzed in our study using the OECD high-growth definition.

Within the scope of the study, we used the following OECD definition of HGFs (equation (1)) for both turnover, employee and asset growth. The OECD typically defines high-growth firms as those with at least 10 employees at the beginning of the growth period and a minimum of annual growth rate of 20% over three consecutive years.

$$\frac{(x_t - x_{t-3})}{x_{t-3}} > 0.728 \;\; if \; x_{t-3} \geq 10 \tag{1}$$

The ratio of firms with an average OECD employee growth between 2012 and 2022 to total firms was 10.01%, while the average turnover growth was 21.19%. In addition, for the eleven years subject to the study, OECD-defined employee growth for non-HG growth firms averages −2.50% while HG is 43.90% and turnover growth for non-HG growth firms averages −1.87% while non-HG is 17.06%. Means and total share of employee and turnover growth are shared as Table A1 in online appendix. Moreover, the number of unique firms for OECD-defined firms and manufacturing firms by year used in our study are presented as Table A2 in online appendix.

3.2. Input variables

The input variables used are chosen based on the literature according to their ability to affect the probability of high growth and enhance the likelihood of persistence. Firm size and age are the most important traits affecting firm growth since the Gibrat law. Audretsch et al. (2004) expressed that firms grow quickly, but after reaching a certain maturity, the growth rate slows down. The relationship between firm growth and export has been extensively studied and, a positive relationship has been observed (Wagner, 2003) because serving more than one market leads to income diversification. Leverage affects firm growth through the external financing channel and many studies see the lack of financial resources as a major obstacle to firm growth (Becchetti & Trovato, 2002). Another factor that affects growth is the fact that firms will grow faster if they have a constant level of cash to pay for their short-term liabilities (Mateev & Anastasov, 2010). Management effectiveness, measured empirically as a profit margin, is an indicator of a firm's ability to survive if prices fall or production costs increase (Aregbeyen, 2012). Besides, profit plays a dominant role in the firm's capacity to access resources, as it simultaneously provides a source of internal financing. Tangibility is important for firms to invest in fixed assets to expand existing markets or enter new markets or products to foster their growth path (Coad & Guenther, 2014). Research and development expenditure is considered a proxy for innovation activities that provide firm growth (Mazzucato & Parris, 2015).

In addition to these variables, we also include firm heterogeneous

variation, such as growth opportunity (Danbolt et al., 2011), which is used as a market-based growth proxy since Tobin Q cannot be calculated for Türkiye, financial slack (Dai & Kittilaksanawong, 2014), which captures the unused debt capacity of firms and cash flow (Brush et al., 2000); which shows borrowers' ability to repay their debts for making more accurate predictions. We also add variables that contain the firm network relationship by using the declaration of purchase and declaration of sale dataset for firm HG estimation. The newly added variables are the variables of how many unique firms a firm buys from and how many unique firms it sells to in a year and the variables of the weighted growth rate of the firms that firms buy from and the weighted growth rate of the firms that firms sell to, which are weighted by the volume of transactions they make (Javorcik, 2004). We also include regional dummies, sector dummies, firm-size type dummies, firm partnership dummies, public-private sector dummies, high-low indebtedness dummies, young-old firm dummies and exporter-non-exporter dummies to analyze details of the factors apart from the balance sheet and profit loss statements. Table A3 online appendix provides descriptive statistics of the data and Table A4 online appendix presents correlation matrix. All models, methods, and independent variables in this study are applied uniformly, with each independent variable measured consistently within the period t.

3.3. Data collection

We construct a unique and comprehensive dataset using various data sources. The primary source is the confidential Credit Registry of the Banks Association of Türkiye, which contains loan-level data encompassing the universe of loan agreements between borrowers and banks from 2009 to 2022. This dataset provides detailed information on loan type, loan amount, issuance date, currency denomination and other relevant details, containing a unique firm ID number. Besides that, the Credit Registry data is linked to the Provisional Income Statements database of all incorporated firms in Türkiye, collected by the Turkish Revenue Administration, which also has the same unique firm ID. This connection enables us to gather critical information on firm size, sector, region and turnover. Additionally, we incorporate firm-to-firm-level transaction data between domestic firms provided by the Ministry of Treasury and Finance.

All databases cover firm-level information merging based on the unique firm number over the 2009–2022 period. Firms not required to report income statements or those that must report under specific conditions are excluded from our analysis due to data constraints. We also cleaned the data and excluded firms with negative assets, turnover and liabilities. The final sample, constructed by integrating all the mentioned micro-databases, our dataset includes 52 percent of total employees, according to the Social Security Institute, allowing us to obtain firm-size information. In addition, based on the Revenue Administration's dataset, our data shows that, on average, 58 percent of total plant, machinery and equipment investments are in Türkiye. Also, the Credit Register database tells us that the stock indebtedness level of the firms included in the study is 53 percent of total indebtedness. We share the definition of all variables in the appendix (see Appendix 1).

3.4. Model specifications and data variations

We employed several models to evaluate the performance of various feature sets and target variables. Our base model (Model 20) uses the target variable OECD Employee HG, which has a binary value (1 for HGF and 0 for Non-HGF) and includes 52 features derived from the full dataset. Model 201 is identical to Model 20 in terms of its features and target variables but is exclusively trained on data from the manufacturing sector. Model 202 shares the same target but excludes the growth opportunity feature. Additionally, we created Model 21 and Model 22, which retain the same 52 features as Model 20 but differ in their target variables, utilizing OECD Turnover HG and OECD Asset HG, respectively (Table 1).

3.5. Data preprocessing

Our analysis involved two fundamental preprocessing steps: addressing class imbalance in our dataset and standardizing feature values. Since HGFs represented only 10% of the total sample, we implemented the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the minority class. By using the auto option from the *Imblearn* package, we effectively balanced the dataset, enabling our models to capture the complexities of high-growth dynamics without biasing towards the majority class. This technique has been proven to improve classification performance (Chawla et al., 2002).

Additionally, we combined SMOTE with random undersampling using an *imbalanced-learn* pipeline (Lemaître et al., 2017). This dual approach was crucial due to the significant class imbalance, facilitating better learning from rare occurrences without merely duplicating data. By leveraging these resampling techniques, we enhanced the model's generalization ability and improved the detection of HGFs while mitigating biases towards the majority class. This strategy reduced the extreme imbalance's impact and allowed for manageable dataset sizes.

It is essential to clarify that SMOTE[1] is used solely within the training phase of the pipeline. During prediction, no synthetic resampling is applied; the model processes only the actual input data. This distinction ensures that our predictions reflect genuine data distributions without any influence from synthetically generated samples.

3.6. Data splitting

To predict HGFs effectively, we adopted a standard data-splitting methodology. We divided our dataset into training, validation and test sets to ensure robust model performance and minimize overfitting, allocating 60% for training, 20% for validation, and 20% for testing (Srhoj, 2022). Our dataset, spanning from 2010 to 2022, was divided into periods: 2010–2013, 2013–2016, 2016–2019 and 2019–2022. This temporal split allows us to evaluate model generalizations across different timeframes, reflecting changes in firm behavior.

While machine learning typically focuses on forward-looking analysis—training, tuning, and testing models with the most extensive available data—we also conducted backward-looking tests. We tested models trained on future periods with earlier data to compare generalization capabilities.

In the first round of analysis, we used the following approach:
Round 1.

1. **Initial Data Split:** We allocated 20% of the first four years of data (2010–2013) for validation and 20% for testing, leaving 60% for model training. This initial split was used only during the first term.
2. **Parameter Tuning:** After splitting the data, the model was trained on the training subset, while the validation set was used to tune the hyperparameters. The goal of the validation split was to optimize parameters such as regularization strength or learning rates without overfitting the test data.
3. **Test and Generalization:** After tuning the parameters, predictions were made on the held-out test data. Fig. 1 illustrates the data split into training, validation and test data.

We extended the evaluation to subsequent periods without retraining

[1] We used SMOTE only in our training and validation stages to avoid overfitting. We also implemented stratified cross-validation to enhance the generalization power of our models and ensure fair computations. We did not use SMOTE in the prediction stages, both within-period and out-of-sample predictions.

Table 1
List of models.

Model Code	Model Name[a]	Target Variables	Features	Data Subset	Notes
Model 20	Base model	OECD employee HG	52 features	All sectors	Base model
Model 201	Filtered model	OECD employee HG	52 features	Manufacturers	Manufacturer data subset
Model 202	Narrowed model	OECD employee HG	51 features	All sectors	Excludes "growth opportunity"
Model 21	Alternative model	OECD turnover HG	52 features	All sectors	Same features as model 20
Model 22	Second alternative model	OECD asset HG	52 features	All sectors	Same features as model 20

[a] In addition to the employee and turnover-defined HG rates in the OECD definition, asset growth was also included in the study. Furthermore, a stratified sector analysis was performed to comprehend how the model's outcomes varied for different groups. Another model was developed by excluding the growth opportunity variable, which exhibited the highest significance. From here on, the models will be referred to by their names as listed in Table 1.

the model, allowing us to assess performance across a large out-of-sample data set.

In the second round (as shown in Fig. 2), we repeated the process, dividing the second period's data into training, validation and test subsets, allowing us to evaluate results while retaining trained models using sci-kit-learn functions and Python's built-in pickle library.

We continued this method for two additional rounds, applying the same protocols illustrated in Figs. 1 and 2. This iterative approach led to training in four distinct rounds. Each round defined new training and validation subsets, emphasizing model evaluation on test data and alternate periods to assess generalization to unseen data.

We repeated these rounds for an additional two periods by applying the same rules as we illustrated in Figs. 1 and 2.

3.7. Cross-validation

During training, we used a 5-fold cross-validation strategy for all models. This process involved splitting the training data into five subsets, with each subset serving as a validation set in rotation. By averaging the performance metrics from each iteration, we ensured that our models were exposed to diverse training and validation data, reducing the likelihood of overfitting and improving generalization capabilities.

3.8. Model parameters summary

We standardized our inputs before training to ensure equal contributions from all features, which is important for data scaling-sensitive models such as Gradient Descent, Logistic Regression (L1), Naive Bayes and KNN. In contrast, tree-based models such as Random Forest and XGBoost inherently handle different feature scales effectively, so no standardization was required. Table A5 online appendix provides model parameters summary. This study contributes to the literature by utilizing a substantial volume of observations and applying seven different algorithms, demonstrating their performance and generalization capabilities. Each model has distinct advantages and disadvantages. Ensemble models such as Random Forest, Gradient Descent, and XGBoost excel in new predictions and generalizations due to their ability to uncover complex relationships. However, they can be difficult to interpret. Conversely, linear models like logistic regression are often slower or less powerful than ensemble models but are easier to interpret (Sahin, 2020; Kabiraj et al., 2020; Kirasich et al., 2018).

Hyperparameter optimization was performed using GridSearchCV to identify the best configurations for each model, incorporating data preprocessing steps such as feature scaling through StandardScaler.

3.9. Robustness checks

To ensure the consistent and accurate performance of our machine learning models, we conducted rigorous unit tests across our functions. By utilizing our abstract structure, we swiftly identified and rectified errors as these functions were applied to independent data across various algorithms. Additionally, we conducted robustness checks through subsetting, evaluating model stability on various random samples from the dataset. This approach allowed us to assess how well our models performed consistently across different data distributions, fostering confidence in the findings.

We further validated our abstract approach through cross-verification. This process involved applying the framework to different

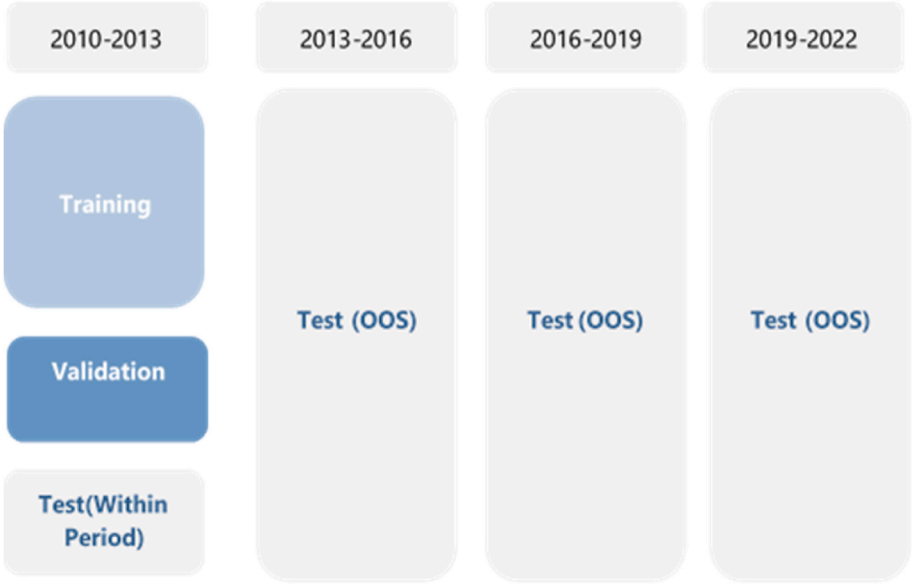


Fig. 1. Round 1 of splitting data as training, validation and testing.

Fig. 2. Round 2 of Splitting Data as Training, Validation and Testing
OOS means Out of sample.

data subsets and comparing the results, reinforcing its framework's reliability and demonstrating its adaptability and effectiveness under varying conditions.

3.10. Evaluation metrics

In predicting HGFs, conventional evaluation metrics like accuracy, precision, recall, F1-score and ROC AUC are widely used (Weinblat, 2018; Megaravalli & Sampagnaro, 2018). While these metrics offer valuable insights, they may not fully capture the challenges posed by the imbalanced datasets typical in HGF studies, where only a small subset of firms achieves rapid growth (Delmar et al., 2003). Therefore, we placed particular emphasis on the Matthews Correlation Coefficient (MCC), as it considers the entire confusion matrix and provides a balanced measure that remains informative even in the presence of significant class disparities (Boughorbel et al., 2017; Matthews, 1975).

Although MCC is not yet the predominant metric in the existing HGF literature, recent studies have highlighted the limitations of conventional metrics in imbalanced contexts. They suggest that MCC provides a more comprehensive evaluation of classifier performance, leading to more reliable insights into the factors driving firm growth (Lundberg & Lee, 2017). This paper aims to contribute to the literature by employing MCC alongside traditional metrics for a robust assessment of model efficacy in predicting HGFs, particularly in imbalanced datasets.

Additional metrics, including accuracy, balanced accuracy, precision, recall, F1-score and confusion matrices, were calculated to provide a comprehensive assessment of each model's performance.

3.11. SHAP values for model interpretation

In addition to evaluating the performance of our HGF classification models, we reported Shapley Additive exPlanations (SHAP) values to interpret each feature's contribution to the model predictions. SHAP values provide a unified measure of feature importance by attributing each feature's contribution to the overall prediction based on game theory principles. This method assists in understanding which features significantly influence HGF predictions and the direction and magnitude of their impact. Moreover, SHAP values enable consistent explanations of feature influence across various model types.

To compute SHAP values, we utilized the methods introduced by Lundberg and Lee (2017), enhancing the interpretability of our machine-learning models. By utilizing SHAP values, we can gain insights into the influence of specific firm characteristics, such as turnover or employee growth, on the likelihood of classification as a high-growth firm. Understanding these underlying data patterns is vital for ensuring the robustness and practical application of our model. We employed the shap and matplotlib Python packages to visualize and report these findings.

4. Results

In this section, we report the performance metrics for seven algorithms, including Logistic Regression with L1 and L2 regularization, Gradient Descent, XGBoosting, Random Forest Classification, KNN and Naive Bayes. We first present the metrics derived from the test data that belong to the same period as the training data. Subsequently, we assess model performance using our trained models to predict outcomes on out-of-sample test data, which comprises 100% of new data from other periods, with no splits. We have shared the details of the models in Table 1 in the methodology section. The evaluation was conducted in two stages: first by testing the models on within-period data (test data from the same period as the training data) and second by assessing the models on out-of-sample data from other periods. Key performance metrics, additional performance metrics, confusion metrics, MCC scores, ROC AUC scores for Model 20, Model 21, Model 201 and Model 202 are given as Tables A6, A7, A8 and A9 in online appendix respectively. These tables provide detailed aggregated results.

4.1. Within-period performance of algorithms (2010–2013)

In the first stage, we tested our models using data from the same period as the training data. This test data represents unseen examples from the training period and is used to evaluate how well the models generalize within the same temporal context. For 2010–2013, we assessed the within-period performance of seven algorithms across three distinct models to determine their effectiveness in predicting high-growth firms. The algorithms included XGBoost, Random Forest, Logistic Regression (L1 and L2 regularization), K-Nearest Neighbors (KNN), Naive Bayes and Gradient Boosting. The within-period results for 2010–2013 emphasize the robustness of tree-based algorithms, particularly XGBoost and Random Forest, in the HGF task, while simpler algorithms like Naive Bayes and KNN consistently showed lower predictive power.

Model 20, which utilized the full dataset with 52 features, demonstrated notable predictive performance across all algorithms, particularly with tree-based methods. XGBoost achieved an ROC AUC of 0.86

and an MCC of 0.37, indicating a high level of predictive power. Random Forest followed closely with an ROC AUC of 0.84 and an MCC of 0.33 (Fig. 3). Logistic Regression with L1 and L2 regularization had an MCC of 0.26 and close ROC AUC values of 0.77, suggesting they performed reasonably well despite being outperformed by tree-based models.

Gradient Descent, on the other hand, achieved an ROC AUC of 0.72 but had an MCC of 0.26, indicating the significance of the MCC in reflecting performance beyond ROC AUC, particularly in the context of our imbalanced dataset.

Fig. 3 presents the average metrics for our base model (Model 20), with XGBoost achieving an average MCC of 0.37, Random Forest at 0.33, and L1, L2, and Gradient Descent models showing MCC values around 0.26. In contrast, Naive Bayes and KNN showed performance levels akin to random guessing, indicating significant limitations in their ability to distinguish between high-growth and non-high-growth firms.

XGBoost demonstrated competitive performance across all evaluation periods, as reflected by its robust metrics: ROC AUC, F1 score and MCC. Its consistent performance underscores XGBoost's strong generalization capabilities, positioning it as a highly competitive option alongside the top-performing model in this study. Notably, although Logistic Regression (L1 and L2) and Gradient Descent did not reach the heights of ensemble methods, they still delivered promising results. This indicates that regularization techniques can enhance accuracy even compared to more complex algorithms. Gradient Descent's performance was enhanced by the inclusion of polynomial features, albeit at the expense of increased computation time.

In contrast, KNN and Naive Bayes exhibited the lowest scores, with MCC values nearing zero. This suggests these models struggled significantly to distinguish between high-growth and non-high-growth firms, reaffirming their unsuitability for this classification problem. The findings suggest that while metrics like ROC AUC and F1 score provide valuable insights, metrics such as MCC provide a deeper understanding of performance, especially in the context of imbalanced datasets. With XGBoost consistently achieving the highest MCC scores, followed closely by Random Forest, the results highlight the superiority of tree-based methods in this specific application, particularly when dealing with imbalanced classes.

In Model 201, we restricted the manufacturer data to evaluate stratified data in the sector and also to see the robustness of our model performance declined slightly due to the narrower data scope. XGBoost remained the top performer with an MCC of 0.32, followed by L1 and L2 regularization (0.29) and Random Forest at 0.28 (see Figure F1 online appendix). Gradient Descent followed with an MCC of 0.20. Although the results were similar, the MCC score allowed for distinguishing their order of performance. These results indicate that with a large amount of data, XGBoost outperforms other models due to its complex nature.

In Model 202, which excluded the "Growth Opportunity" feature to see if growth opportunity drove the outcome metrics, there was a slight decrease across all algorithms except XGBoost, illustrating the importance of this feature. XGBoost consistently demonstrated an MCC of 0.28 and a ROC AUC of 0.82, while Random Forest followed with an MCC of 0.20 (see Figure F2 online appendix). However, L1 and L2 regularization had an MCC of 0.19, suggesting that complex algorithms like XGBoost manage the omission of important variables better than simpler alternatives.

Fig. 4 illustrates the ROC curves for all seven algorithms for our base model. To assess the generalization power of our algorithms, we saved our training models using the technique discussed in the methodology section. We employed these trained models to predict a new period of data without further training. The results of out-of-sample testing are presented in the following subsection.

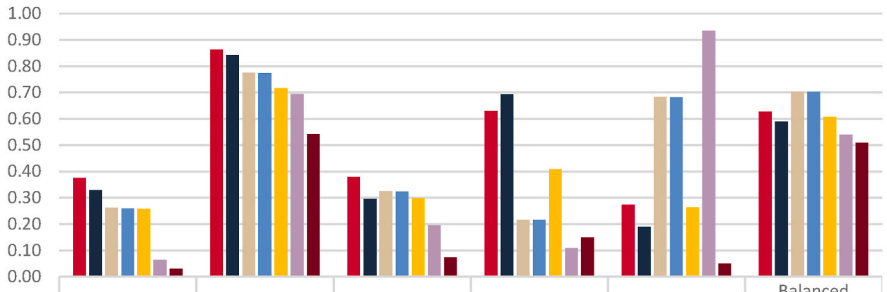
4.2. Out-of-sample performance

Our aim is to identify models with the highest generalization power, rather than those that perform well in only a single or a few periods. In predictive modeling, especially HGF identification, model stability often outweighs the need for peak performance on a particular dataset. A consistent and moderately performing model is generally preferable because it provides reliable and repeatable results, ensuring robustness when applied to unseen data.

In this section, we highlight the generalization capabilities of our models. The trained models were applied to new period data to evaluate their predictive power in an out-of-sample context. The results of the out-of-sample tests largely mirrored those of the within-period tests, with XGBoost and Random Forest again demonstrating superior performance. Both algorithms maintained high MCC and ROC AUC scores, indicating their ability to retain predictive ability when exposed to data from different periods.

4.3. Base model performance

The results of out-of-sample testing mostly mirrored those of the within-period test. XGBoost and Random Forest demonstrated superior performance, maintaining high MCC and ROC AUC scores, indicating that these models retained their predictive capabilities even when



	MCC	ROC AUC	F1	Precision	Recall	Balanced Accuracy
XGBoost	0.37	0.86	0.38	0.63	0.27	0.63
Random Forest Classifier	0.33	0.84	0.30	0.69	0.19	0.59
Logistic Regression with L1 regularization	0.26	0.77	0.33	0.22	0.68	0.70
Logistic Regression with L2 regularization	0.26	0.77	0.32	0.22	0.68	0.70
Gradient Descent	0.26	0.72	0.30	0.41	0.26	0.61
Naive Bayes	0.06	0.69	0.19	0.11	0.93	0.54
KNN	0.03	0.54	0.07	0.15	0.05	0.51

Fig. 3. Base model (Model 20) – within period average metrics.

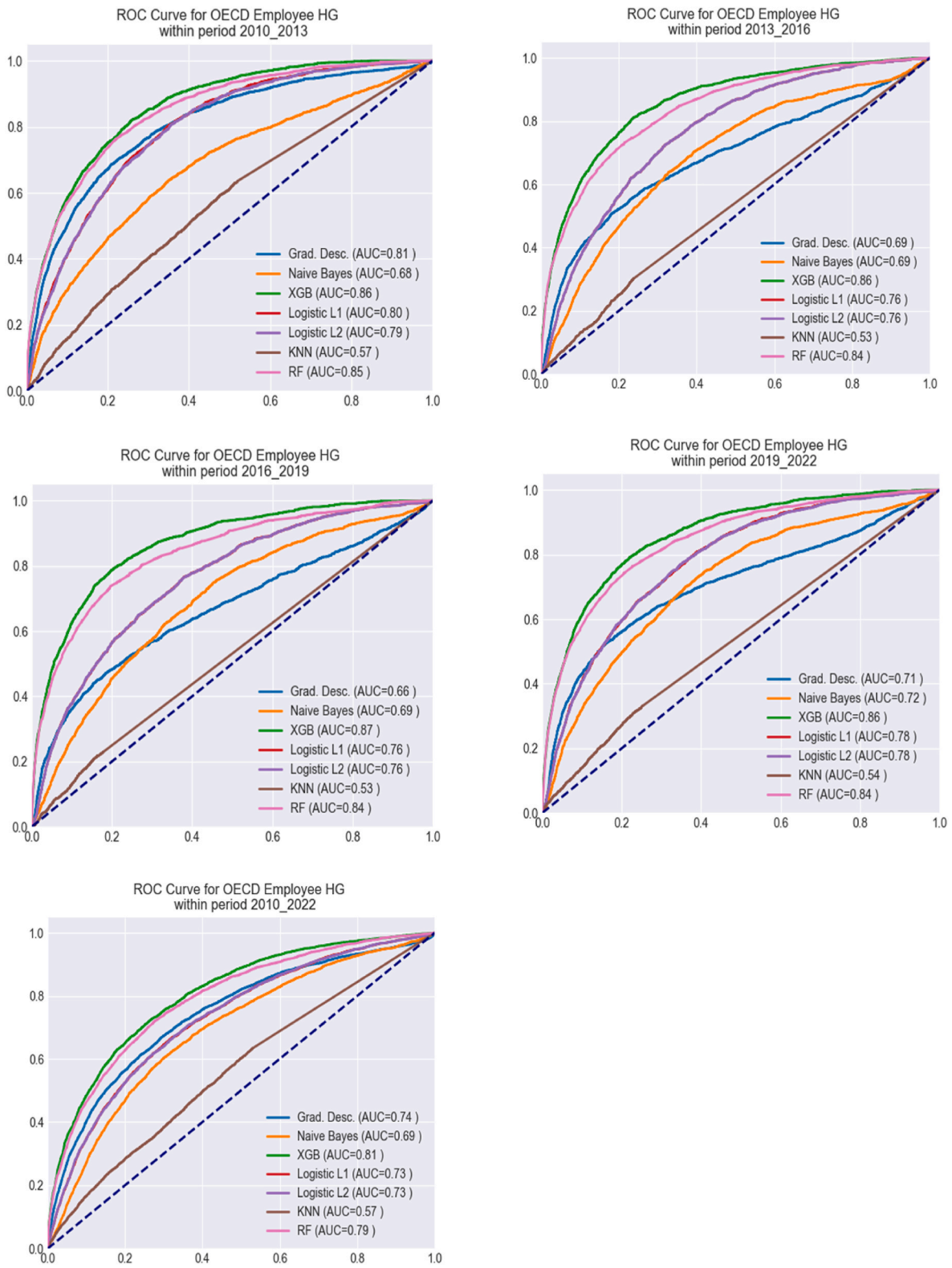
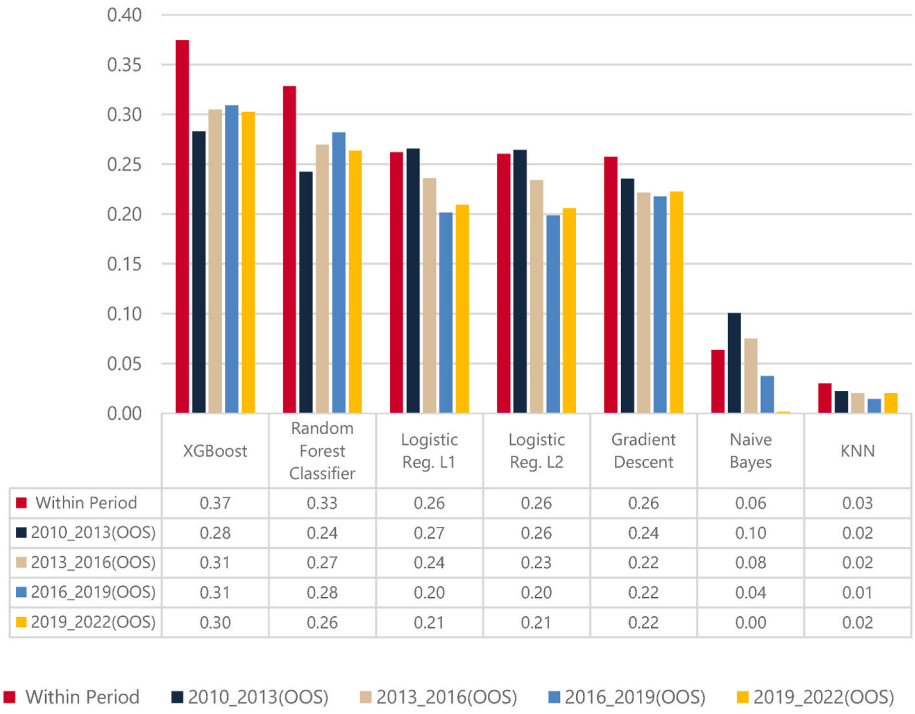


Fig. 4. Roc curves for the base model – within period test data.

exposed to data from different periods. As shown in Fig. 5, XGBoost achieved an average MCC score of 0.30 across all periods, while Random Forest followed closely with an average MCC score of 0.26. Both models consistently achieved ROC AUC scores above 0.75, highlighting their strong generalization capabilities across varying time horizons. Logistic Regression with L1 regularization and L2 regularization also performed well in the out-of-sample tests, with an average MCC score of 0.23, notably higher than that of other non-ensemble methods.

These findings further validate the superior capability of ensemble methods, such as XGBoost and Random Forest, in capturing the complex, non-linear relationships within the dataset, which are crucial for predicting high-growth firms. The consistent performance of these



	XGBoost	Random Forest Classifier	Logistic Reg. L1	Logistic Reg. L2	Gradient Descent	Naive Bayes	KNN
Within Period	0.37	0.33	0.26	0.26	0.26	0.06	0.03
2010_2013(OOS)	0.28	0.24	0.27	0.26	0.24	0.10	0.02
2013_2016(OOS)	0.31	0.27	0.24	0.23	0.22	0.08	0.02
2016_2019(OOS)	0.31	0.28	0.20	0.20	0.22	0.04	0.01
2019_2022(OOS)	0.30	0.26	0.21	0.21	0.22	0.00	0.02

Fig. 5. MCC scores for Base Model.

models across different periods suggests an adaptability to changes in firm dynamics, enabling reliable predictions in real-world settings. Conversely, KNN and Naive Bayes continued to underperform in out-of-sample testing, yielding MCC scores close to zero. This indicates their limited capacity to handle temporal shifts and complex feature interactions, underscoring their unsuitability for this classification problem. We also share ROC AUC scores and MCC scores for the Base model as Figure F3 and Figure F4 in the online appendix.

Figure F5 available at the online appendix summarizes the average MCC scores of the algorithms for the Filtered Model. This model retains the same dependent variables and features as the base model but is limited to manufacturer-only data for training and prediction. Because of the reduced number of observations, these models exhibited slightly diminished prediction and generalization power compared to the base model.

It is noteworthy that XGBoost demonstrated superior prediction and generalization power compared to Random Forest and both L1 and L2 models. Conversely, Gradient Descent showed the most considerable decline in performance among the successful algorithms. Our findings imply that XGBoost remains the most robust option for prediction and generalization, particularly when sufficient observations are present. However, while predicting out-of-sample data, especially during Term 1, both Random Forest and XGBoost encountered challenges in accuracy compared to L1 and L2 models.

Table A10 online appendix presents the coefficients of L1 and L2 for the Narrowed Model (Model 202), which excludes the important "Growth Opportunity" feature from the base model. While the predictive capability decreased as previously highlighted, this model continues to identify key features with reasonable performance.

These coefficients highlight the importance of specific variables throughout the periods studied. In particular, while the removal of the Growth Opportunity feature results in a decline in predictive performance, the remaining features still offer critical insights into the factors associated with high growth. The insights gained from these coefficients can inform practitioners and researchers on the most influential predictors of firm growth, guiding future strategies and research in the domain.

4.4. Feature selection and model insights

This paper aims to evaluate several machine learning algorithms for predicting HGFs, with a dual focus on prediction accuracy and model interpretability. Specifically, we aim to compare the performance of XGBoost, Random Forest, and Logistic Regression with L1 regularization, among others. Each of these models offers unique advantages: XGBoost and Random Forest excel in predictive power due to their ability to capture complex, non-linear interactions among features, while Logistic Regression with L1 regularization prioritizes interpretability by selecting a sparse subset of features that are most influential for prediction.

Through a thorough comparison, we aim to quantify how much predictive performance is sacrificed for greater interpretability. For instance, XGBoost achieved an MCC of 0.30 in out-of-sample predictions, significantly outperforming L1 with an MCC of 0.23. However, the L1 model provides clearer insights into the key factors that drive firm growth. This trade-off is crucial in contexts where stakeholders prioritize both accuracy and transparency. Consequently, this study bridges a gap in high-growth firm literature by addressing the need for predictive performance and model interpretability.

A secondary aim of this study is to retrieve the most effective features in predicting HGF. Using L1 regularization, we can identify the features that contribute most to firm growth. L1 regularization reduces some coefficients to zero, resulting in a sparse model that emphasizes the most relevant variables. This approach contrasts with tree-based models like XGBoost, where feature importance is harder to interpret due to the model's complexity. While XGBoost provides excellent predictive performance, Logistic Regression with L1 is more interpretable and offers actionable insights regarding feature importance.

Previous research in the HGF literature has largely focused on economic and firm-level factors affecting growth but often neglected the use of machine learning models for robust feature selection. By applying L1 regularization, this study identifies key drivers of firm growth, allowing us to understand the factors that contribute to the rapid expansion of firms. These insights are not only valuable for academic purposes but also for policymakers and investors who seek to identify high-growth

firms early and allocate resources effectively.

4.5. Coefficients of L1 and L2

In this section, we present the coefficients obtained from our successful linear models, L1 and L2, for the base model across all periods. These algorithms hold particular significance due to their linear nature, with coefficients illustrating the direction and magnitude of features in predicting high-growth firms.

For the Base Model, Table 2 displays the first 10 coefficients derived from L1 and L2 algorithms. The L1 (Lasso) regression, which performs feature selection by driving certain coefficients of variables to zero, exhibited predictive accuracy and generalization similar to L2 (Ridge) regularization, which tends to retain all features. Key features consistently appearing in the top ten include Growth Opportunity (+), Medium Firms dummy (+), Firm Age (−) and Size (+).

4.6. Feature importances

Table 3 provides a detailed overview of the important features derived from the various models used in our study. Our supplementary section includes feature importance tables for all periods. Notably, our top-performing four models consistently list Growth Opportunity among the first four most important features, except the Gradient Descent algorithm. The Medium Firms dummy also appears prominently across all models.

Logistic Regression with both L1 and L2 regularization demonstrates a similar ordering of features in this term. While slight differences exist in feature ranking, this similarity suggests that L1 regularization effectively highlights critical features while maintaining comparable performance with L2, which retains all features.

4.7. SHAP (Shapley Additive exPlanations) analysis

SHAP values provide a unified measure of feature importance by distributing each feature's impact on the prediction across all possible feature combinations. Drawing from concepts in cooperative game theory, specifically Shapley values, it equitably assigns contributions among features based on their participation in multiple coalitions.

One of SHAP's key advantages is its alignment with local interpretability, as it assigns an importance score to each feature for individual predictions. These scores can be visualized in SHAP summary plots, showing the distribution and direction of each feature's impact on the model's predictions. Using SHAP, we can determine that features with higher absolute SHAP values have a greater influence on predictions, while the sum of all SHAP values for a specific observation corresponds to the model's output.

In this study, we employed SHAP to enhance the interpretability of our complex models, particularly XGBoost and Random Forest. Utilizing Python packages, sci-kit-learn and shap, we generated SHAP summary plots that communicate the prominent features instrumental in predicting high-growth firms (Fig. 6).

4.8. Summary of model performances

Our findings highlight the effectiveness of machine learning models, especially XGBoost and Random Forest, in classifying high-growth firms. Careful feature selection and thorough model evaluation are crucial for achieving accurate and interpretable results. Tree-based algorithms demonstrate superior performance in classifying high-growth firms, particularly when dealing with imbalanced datasets.

5. Discussion

This study provides a comprehensive evaluation of machine learning

Table 2
Coefficients of L1 and L2 for base model (model 20).

	2010–2013		2013–2016		2016–2019		2019–2022	
L1	Feature	Coef.			Feature	Coef.		
	Growth opportunity	0.99	R&D expenditures/asset (million)	−13.56	Growth opportunity	0.73	Growth opportunity	0.70
	Medium firms dummy	0.58	Growth opportunity	0.73	Firm age	−0.41	Firm age	−0.49
	Small firms dummy	0.42	Limited firms	0.57	Firm number in firm seller's network	−0.34	Size	0.35
	Size	0.41	Incorporated firms	0.45	Size	0.31	Construction sector dummy	0.24
	Firm age	−0.27	Firm age	−0.40	Manufacturing sector dummy	−0.28	Medium firms dummy	0.20
	Productivity	−0.20	Size	0.31	Medium firms dummy	0.24	Service sector dummy	0.18
	Gross profit profitability	0.17	Small firms dummy	−0.26	Marmara region dummy	−0.23	Weighted averaged employee in firm buyer's network	0.17
	Limited firms	0.15	Construction sector dummy	0.23	Trade sector dummy	−0.21	Total Debt(million)	0.14
	Construction sector dummy	0.13	Service sector dummy	0.17	Firm number in firm buyer's network	0.19	High debt firm dummy	−0.14
	Gross profit margin	−0.13	Long term debt(million)	−0.15	Central Anatolia region dummy	−0.14	Gross profit profitability	0.10
L2	2010–2013		2013–2016		2016–2019		2019–2022	
	Growth opportunity	1.03	Growth opportunity	0.73	Growth opportunity	0.73	Limited firms	3.04
	Medium firms dummy	0.71	Firm age	−0.40	Firm age	−0.41	Incorporated firms	2.91
	Small firms dummy	0.60	Limited firms	0.34	R&D expenditures/asset (million)	−0.37	Growth opportunity	0.72
	Size	0.48	Size	0.31	Firm number in firm seller's network	−0.35	Firm age	−0.54
	Firm age	−0.36	Small firms dummy	−0.26	Size	0.32	Cooperatives	0.39
	Productivity	−0.27	Construction sector dummy	0.23	Manufacturing sector dummy	−0.29	Size	0.37
	Gross profit profitability	0.22	Incorporated firms	0.22	Marmara region dummy	−0.25	Others (other than Ltd., Inc and Coop. firms)	0.37
	Cash flow	0.20	Service sector dummy	0.17	Limited firms	0.25	Construction sector dummy	0.29
	Gross profit margin	−0.19	Long term debt(million)	−0.15	Medium firms dummy	0.24	Service sector dummy	0.25
	Trade sector dummy	−0.16	Medium firms dummy	0.14	Trade sector dummy	−0.22	Medium firms dummy	0.23

Table 3
List of feature importances for 2010–2013.

Gradient Descent
R&D expenditures/assets (million)^2
Number of branches short term debt (million)
Number of branches profit or losses before taxes (million)
Growth opportunity
Medium firms dummy
Logistic Reg. L1
Growth opportunity
Medium firms dummy
Small firms dummy
Size
Firm age
Logistic Reg. L2
Growth opportunity
Medium firms dummy
Small firms dummy
Size
Firm age
Random forest classifier
Growth opportunity
Medium firms dummy
Small firms dummy
Construction sector dummy
Firm number in firm buyer's network
XGBoost
Growth opportunity
Medium firms dummy
Productivity
Cash flow
Small firms dummy

algorithms for predicting HGFs, addressing the limitations of previous research by utilizing a large dataset, multiple evaluation metrics, and a rigorous cross-validation strategy. Ensemble methods, like XGBoost and Random Forest, exhibit superior predictive performance. The consistent outperformance of these tree-based models, even in out-of-sample testing, demonstrates their robustness in capturing the complex, non-linear firm growth dynamics.

The predictive performance of Logistic Regression with L1 regularization is notable, providing accuracy comparable to ensemble methods while offering greater interpretability through feature selection. This contrasts with the less interpretable but accurate XGBoost and Random Forest models, showing a trade-off between predictive power and model transparency. Conversely, the significantly lower performance of KNN and Naive Bayes, with MCC scores approaching zero, underscores the challenges simpler algorithms face in capturing the complexities of HGF prediction. Additionally, the relatively better performance of L2-regularized Logistic Regression compared to KNN and Naive Bayes further emphasizes the advantages of regularization techniques in enhancing model accuracy.

Our analysis reveals that while XGBoost initially exhibited superiority in generalization power compared to other algorithms, this advantage diminished in subsequent rounds. Over time, the differences in generalization performance between XGBoost and other models, such as Random Forest and Logistic Regression, became less pronounced, suggesting that XGBoost's forward-looking generalization power may depend on specific data characteristics or time-specific factors.

In other words, ensemble methods, especially XGBoost, demonstrate strong generalization to future periods (see section 4.2). Other algorithms such as RF, GD, L1 and L2 showed moderate generalization power in both future periods and earlier periods' predictions.

A more valid, real-life approach would be to use the biggest data available to train and predict with new data in order to maximize the learning possibilities of the algorithms or to check expanding windows or rolling windows to see the prediction power of these models and algorithms. However, we left them for future works concentrating on this study to see the differences in prediction powers and generalizations when training data is limited to only one period. In addition, we thought having a similar and smaller size of observations would make comparisons easier and force our algorithms to learn and adapt to different circumstances.

In our analysis, we found that when focusing on manufacturers, complex models like XGBoost remained the top performers. However, as the number of observations decreased, the performance gap between complex models and simpler ones, such as logistic regression, narrowed. This trend is consistent with prior research. Caruana and Niculescu-Mizil (2006) observed that the advantages of complex models diminish with smaller training sets, as simpler models are less prone to overfitting. Similarly, Fernández-Delgado et al. (2014) noted that the performance difference between complex and simple models shrinks when data size is reduced, making simpler models more competitive. Our findings align with these studies, emphasizing that dataset size plays a critical role in model performance.

Our feature importance analysis (Table 3) reveals the consistent significance of certain factors across various models. Notably, Growth Opportunity consistently ranks highly, underscoring the critical role of market conditions and firm-specific potential for future growth. The frequent appearance of the Medium Firm Dummy as a key feature further supports this interpretation. The notable similarity in feature rankings between L1 and L2 regularized Logistic Regression suggests that the feature selection process inherent in L1 regularization does not significantly compromise predictive accuracy (Table 3). For a more comprehensive understanding of feature influence, SHAP analysis (Lundberg & Lee, 2017) provides nuanced insights into individual predictions, detailing how various features contribute to overall model outputs (see Figure F6 online appendix).

Our analysis of SHAP values from the ensemble models revealed that

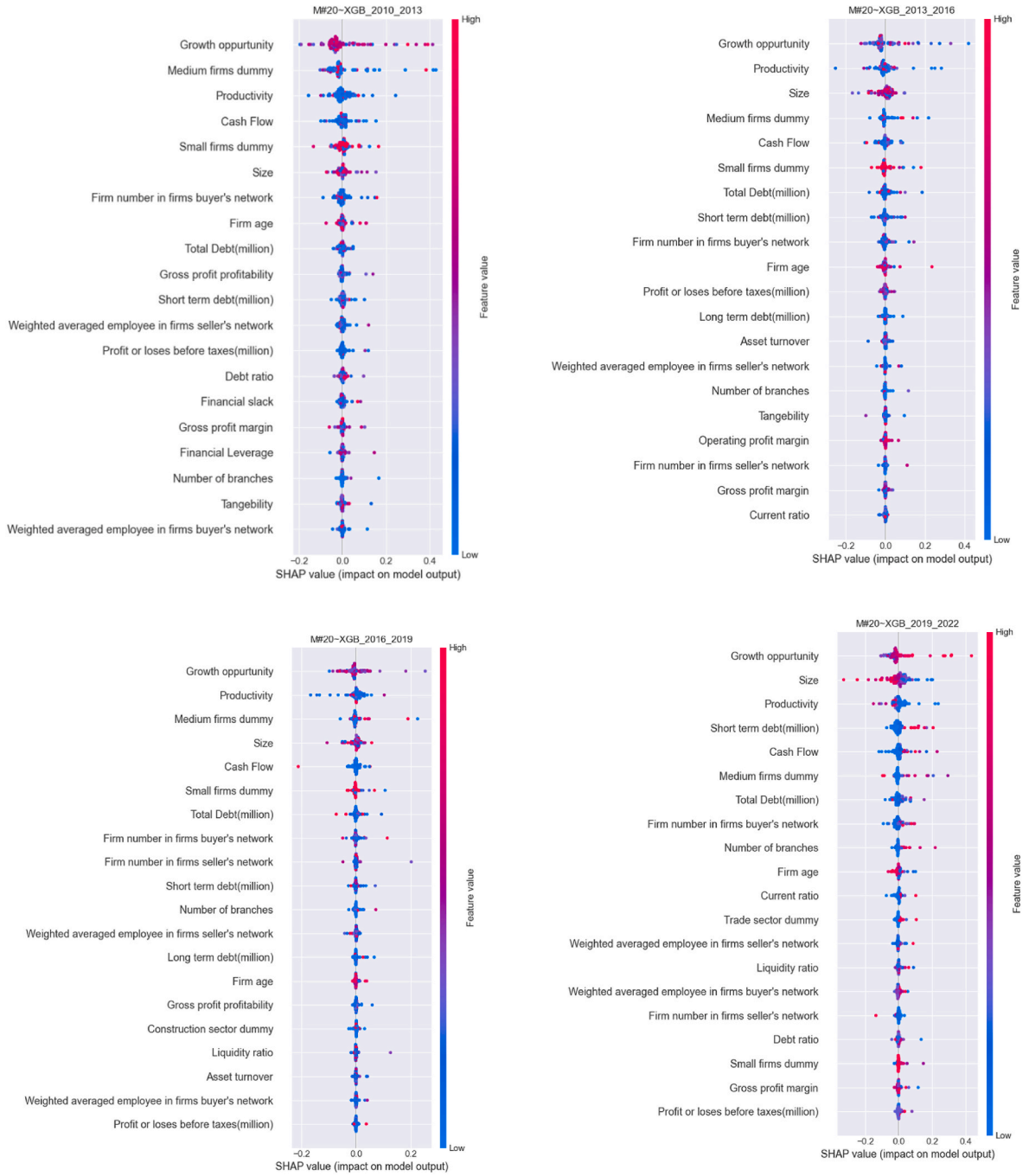


Fig. 6. Shap summary plot for base model (XGBoost).

the most important features identified by these models were consistent with the key features highlighted by the linear models.

Upon analyzing the feature importances, we found that firm age (Weinblatt, 2018; Houle & Macdonald, 2023) and firm size (Weinblatt, 2018; VanWitteloostuijn & Kolkman, 2019), which have been extensively studied in the literature, significantly impact firm growth. Throughout all periods, firm age, defined as the difference between the date of incorporation and the date of the balance sheet publication, hurts growth. Conversely, firm size, represented as the logarithm of total assets, consistently increases the probability of high-growth firms.

One reason for the decrease in the probability of achieving HGF as firm age increases is supported by life cycle theory, which posits that firms experience rapid growth until they reach maturity, at which point growth slows due to market saturation. This perspective suggests that as firms age, their growth opportunities diminish, highlighting the importance of growth potential. Additionally, over time, firms may become more risk-averse, hindering their ability to achieve significant growth through innovative products and leading to diminishing returns on growth (Coad et al., 2013). In contrast, larger firms are more likely to become HGFs (Bottazzi & Secchi, 2006), as they benefit from economies of scale that reduce costs, gain easier access to financial resources, and can expand their reach by penetrating international markets.

Access to internal and external resources is crucial in determining whether a firm can achieve high growth, and this impact varies based on several factors, such as a firm's growth stage, the industry it operates in and prevailing economic conditions. Our analysis indicates that while a firm's external funding sources, encompassing both short- and long-term borrowing, positively influence growth, their effect on high-growth

firms (HGFs) is significantly greater than that of internal cash flow. This finding highlights a vital lesson: for firms in developing countries like Türkiye, leveraging external resources is essential to fuel growth.

Interestingly, even though we see that internal resources have a considerable impact—depending on the specific period and methodology used—cash on hand does not prove to be as vital as accessing a blend of both internal and external financing options. This distinction is critical for practitioners and policymakers alike. Furthermore, we observe that high levels of indebtedness consistently exhibits a negative correlation with growth across all analyzed periods. This suggests a nuanced perspective, where borrowing can support a firm's journey toward becoming a HGF, but only up to a certain threshold. Beyond that point, excessive debt can hinder growth, a reminder of the delicate balance firms must maintain in managing their financial leverage.

The number of firms a firm engages with, both as buyers and sellers, along with the growth of those trading partners—weighted by trade volume—emerges as a significant factor in increasing the probability of becoming a HGF. This network effect supports the development of expertise, providing firms access to valuable resources. Long-term customer relationships, in particular, create an environment of trust, which increases the likelihood of firms achieving high growth rates as order volumes expand alongside the growth of their trading partners.

Moreover, relationships with buyers encourage collaborative innovation and enhance product development efforts (Ramaswamy & Ozcan, 2016). Simultaneously, relationships with suppliers mitigate risks by enabling access to diversified supply chains (Christopher, 2016). This interconnectedness is not just beneficial but essential, as our findings illustrate. Particularly in the context of the Gradient Descent method, the network variable consistently ranks among the top five influencing factors across different periods. As firms navigate the complexities of the economy, those that adeptly manage their buying and selling relationships—especially within a heterogeneous firm landscape—stand out even more.

Overall, our findings suggest that while ensemble models like XGBoost and Random Forest provide superior predictive accuracy, linear models with L1 regularization offer valuable interpretability without significantly compromising performance. The alignment of feature importance across both model types further supports the reliability of our conclusions. For practical applications, such as guiding investment decisions or shaping policy, a combination of both model types may be the most effective approach—leveraging the accuracy of ensemble models and the interpretability of linear models to provide a comprehensive understanding of the factors driving high-growth firms.

5.1. Future work

Building on our study of classifying high-growth firms in Türkiye using seven different algorithms, future research could explore the integration of advanced machine learning techniques such as deep learning or ensemble methods to enhance predictive accuracy. Investigating the nuances of different algorithms across distinct industry sectors could enable a more tailored approach to identifying high-growth firms. Future research could explore the integration of additional firm-specific and macroeconomic variables to enhance the predictive power of these models. Future work could also focus on longitudinal studies that track the evolution of firm performance over time, assessing how shifts in market dynamics or regulatory environments influence HGF classification outcomes. Lastly, incorporating alternative data sources, such as textual data from company reports or social media sentiment analysis, could enhance prediction accuracy and provide additional insights into the factors driving HGF growth.

5.2. Limitations of the study

While our study makes valuable contributions, several limitations should be acknowledged. Although we rigorously tested our algorithms against multiple unseen out-of-sample datasets and found consistent results within-period test data, this approach may still underrepresent certain variable interactions or external factors influencing high growth. The focus on Türkiye might also limit the applicability of our findings to firms in different cultural or economic contexts, potentially overlooking regional variations in growth drivers. Furthermore, the algorithms employed may vary in interpretability, which could pose challenges when applying insights derived from the models in practical business settings. Addressing these limitations in future research could further substantiate the findings and enhance their applicability.

One noteworthy finding in our analysis is that not all periods provided equally beneficial performance for the algorithms regarding their generalization capabilities. We discovered that Gradient Descent, along with our tree-based models, XGBoost and Random Forest, did not perform as effectively when trained on data from Term 4 (2019–2022) compared to earlier periods. In contrast, the Logistic Regression models (L1 and L2) faced challenges during 2010–2013, experiencing a decline in generalization power. While all models have limitations, it is important to note that some models prove more useful than others, particularly when their training periods provide a good representation of the conditions in the data they are applied to.

The COVID-19 pandemic further complicates our analysis of the last period. The unprecedented economic shifts and disruptions caused by the pandemic inherently make it more challenging for algorithms to identify general patterns from this period when predicting outcomes for other timeframes. Insights garnered from earlier periods may not seamlessly translate to the unique context created by the pandemic, highlighting the complexities found in real-world data.

6. Conclusion

In this study, we evaluated the performance of several machine learning models in predicting high-growth firms, using a comprehensive dataset of firm-level characteristics. Our results indicate that ensemble models, such as XGBoost and Random Forest, deliver the highest predictive performance, both within and across periods. However, linear models like Logistic Regression with L1 regularization also performed well, offering a balance between accuracy and interpretability.

The ability to interpret model outputs is particularly crucial in economic research and decision-making. The consistency between the feature importance derived from linear models and the SHAP values of ensemble models underscores the robustness of our findings and highlights the key drivers of firm growth, such as firm size, age, leverage and growth opportunity.

Our results indicate that variations in macroeconomic conditions can influence the prominence of certain variables, such as the construction dummy, in explaining HGFs across different periods. This finding emphasizes the necessity of examining macroeconomic factors alongside micro-level data, as their interplay can significantly affect the dynamics of firm growth.

In conclusion, this study provides critical insights into the application of machine learning for HGF prediction, highlighting the superior performance of ensemble methods, the importance of feature selection, and the necessity of incorporating multiple evaluation metrics, particularly in imbalanced datasets. In our study, ROC AUC values ranged from 0.53 to 0.87 for employee-high growth and from 0.53 to 0.91 for turnover high-growth, depending on the method used. The results underscore the need for a comprehensive approach that prioritizes predictive accuracy and model interpretability, offering practical implications for investors, policymakers, and researchers alike. The limitations and future research directions outlined here provide a roadmap for continued exploration in this important field. As George E. P. Box famously stated, "All models are wrong, but some are useful," reinforcing that while no model is perfect, certain models can provide invaluable insights when trained on representative data.

In our study, we found that some algorithms performed efficiently

and accurately, such as XGBoost, which demonstrated high predictive capability despite being less interpretable. Others, like Logistic Regression with L1 and L2 regularization, have balanced interpretability with moderate performance. Gradient Descent showed reasonable performance but was computationally intensive, while algorithms like KNN and Naïve Bayes exhibited limited predictive accuracy in our analysis.

Declaration of competing interest

There is no conflict of interest.

Acknowledgements:

The views expressed in this study are those of the authors and do not necessarily represent the official views of the Central Bank of the Republic of Turkey.

Appendix 1. Definition of Variables

Variable	Definition
Size	Log of total assets
Financial	Total liability divided by total assets
Debt ratio	Financial liabilities plus trade debts divided by total assets
Equity ratio	Shareholder's equity divided by total assets
Current ratio	Current assets divided by current liabilities
Cash ratio	Cash and cash equivalents divided by current liabilities
Liquidity ratio	Cash and cash equivalents divided by total assets
Operating profitability	Operating profit divided by total assets
Gross profit profitability	Gross profit divided by total assets
Operating profit margin	Operating profit divided by net sales
Gross profit margin	Gross profit divided by net sales
Firm age	Difference in current year and foundation year
Total debt (million)	Short term and long term liability
Number of branches	Number of branches
Short term debt (million)	Short term debt (million)
Long term debt (million)	Long term debt (million)
Profit or losses before taxes (million)	Profit or losses before taxes (million)
Real R&D exp/employees (million)	Real R&D exp/employees (million)
R&D expenditures/asset (million)	R&D expenditures/assets (million)
Tangebility	Tangible fixed assets divided by total assets
Productivity	Net sales divided by the number of employees
Maturity	Long term debt divided by short term and long term liability
Cash flow	Earnings before interest, tax and depreciation divided by total assets
Financial slack	Fixed asset divided by long term and short term liability
Asset turnover	Net sales divided by total assets
Growth opportunity	Logarithmic change in real net sales deflated by consumer price Index (CPI)
Seller's network firm number	Total number of unique firms buy from the firm in a year
Buyer's network firm number	Total number of unique firms sell to the firm in a year
Seller's network firm growth	Growth of firms sold by the firm within a year, weighted by amount
Buyer's network firm growth	Growth of firms bought by the firm within a year, weighted by amount

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bir.2024.12.001.

References

Acs, Z., Parsons, J., & Tracy, W. (2008). *High-impact firms: Gazelles revisited*. Washington, DC: SBA Office of Advocacy.
Aregbeyen, O. (2012). The determinants of firm growth in Nigeria. *Pakistan Journal of Applied Economics, 22*(1 & 2), 19–38.
Audretsch, D., & Dohse, D. (2007). Location: A neglected determinant of firm growth. *Review of World Economics, 143*, 79–107.
Audretsch, D. B., Klomp, L., Santarelli, E., & Thurik, A. R. (2004). Gibrat's Law: Are the services different? *Review of Industrial Organization, 24*(3), 301–324.
Becchetti, L., & Trovato, G. (2002). The determinants of growth for small and medium-sized firms. The role of the availability of external finance. *Small Business Economics, 19*(4), 291–306.
Bottazzi, G., & Secchi, A. (2006). Explaining the distribution of firm growth rates. *The RAND Journal of Economics, 37*(2), 235–256.
Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One, 12*(6), Article e0177678.
Brush, T. H., Bromiley, P., & Hendrickx, M. (2000). The free cash flow hypothesis for sales growth and firm performance. *Strategic Management Journal, 21*(4), 455–472.
Capon, N., Farley, J. U., & Hoenig, S. (1990). Determinants of financial performance: A meta-analysis. *Management Science, 36*(10), 1143–1159.
Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *In proceedings of the 23rd international conference on machine learning* (pp. 161–168).
Chae, H. C. (2024). In search of gazelles: Machine learning prediction for Korean high-growth firms. *Small Business Economics, 62*(1), 243–284.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
Christopher, M. (2016). *Logistics and supply chain management: Logistics & supply chain management*. UK: Pearson.
Coad, A. (2009). *The growth of firms: A survey of theories and empirical evidence*. Cheltenham: Edward Elgar Publishing.
Coad, A., Daunfeldt, S. O., Hölzl, W., Johansson, D., & Nightingale, P. (2014). High-growth firms: Introduction to the special section. *Industrial and Corporate Change, 23*(1), 91–112.
Coad, A., Frankish, J., Roberts, R. G., & Storey, D. J. (2013). Growth paths and survival chances: An application of gambler's ruin theory. *Journal of Business Venturing, 28*(5), 615–632.
Coad, A., & Guenther, C. (2014). Processes of firm growth and diversification: Theory and evidence. *Small Business Economics, 43*, 857–871.
Coad, A., & Srhoj, S. (2020). Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms. *Small Business Economics, 55*(3), 541–565.

Dai, W., & Kittilaksanawong, W. (2014). How are different slack resources translated into firm growth? Evidence from China. *International Business Research, 7*(2), 1.
Danbolt, J., Hirst, I. R., & Jones, E. (2011). The growth companies puzzle: Can growth opportunities measures predict firm growth? *The European Journal of Finance, 17*(1), 1–25.
Daunfeldt, S. O., & Halvarsson, D. (2014). Are high-growth firms one-hit wonders? Evidence from Sweden. *Small Business Economics, 44*, 361–383.
Davidsson, P., Achtenhagen, L., & Naldi, L. (2010). *Small firm growth. Foundations and Trends® in Entrepreneurship, 6*(2), 69–166. https://doi.org/10.1561/0300000029
Delmar, F., Davidsson, P., & Gartner, W. B. (2003). Arriving at the high-growth firm. *Journal of Business Venturing, 18*(2), 189–216.
Eurostat-OECD. (2008). *Eurostat-OECD manual on business demography statistics*. Paris, France: OECD Publishing.
Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*(1), 3133–3181.
Geroski, P. A., Mata, J., & Portugal, P. (2010). Founding conditions and the survival of new firms. *Strategic Management Journal, 31*(5), 510–529.
Gibrat, R. (1931). *Les inégalités économiques*. Paris: Recueil Sirey.
Gupta, P. D., Guha, S., & Krishnaswami, S. S. (2013). Firm growth and its determinants. *Journal of innovation and entrepreneurship, 2*, 1–14.
Guzman, J., & Stern, S. (2020). The state of American entrepreneurship: New estimates of the quantity and quality of entrepreneurship for 32 US States, 1988–2014. *American Economic Journal: Economic Policy, 12*(4), 212–243.
Henrekson, M., & Johansson, D. (2010). Gazelles as job creators: A survey and interpretation of the evidence. *Small Business Economics, 35*, 227–244.
Hölzl, W. (2014). Persistence, survival, and growth: A closer look at 20 years of fast-growing firms in Austria. *Industrial and Corporate Change, 23*(1), 199–231.
Houle, S., & Macdonald, R. (2023). *Identifying nascent high-growth firms using machine learning*. Bank of Canada Staff Working Paper (No. 2023-53).
Javorcik, B. S. (2004). Does foreign direct investment increase the productivity of domestic firms? In search of spillovers through backward linkages. *The American Economic Review, 94*(3), 605–627.
Kabiraj, S., Kling, N., & Siddik, M. N. A. (2020). Diabetes control in China: Building a supervised machine learning diabetes predictor based on living circumstances of Chinese citizens aged 45 and above. *International Journal of Business and Information Technology, 13*(2).
Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review, 1*(3), 9.
Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research, 18*(17), 1–5.
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems* (p. 4765).
Marsili, O. (2001). The anatomy and evolution of industries: Technological change and industrial dynamics. In *In the anatomy and evolution of industries*. Edward Elgar Publishing.
Mason, C., & Brown, R. (2012). Creating good public policy to support high-growth firms. *Small Business Economics, 40*(2), 211–225. https://doi.org/10.1007/s11187-011-9369-9
Mateev, M., & Anastasov, Y. (2010). Determinants of small and medium-sized fast growing enterprises in central and eastern europe: A panel data analysis. *Financial Theory and Practice, 34*(3), 269–295.
Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure, 405*(2), 442–451.
Mazzucato, M., & Parris, S. (2015). High-growth firms in changing competitive environments: The US pharmaceutical industry (1963 to 2002). *Small Business Economics, 44*(1), 145–170.
Megaravalli, A. V., & Sampagnaro, G. (2018). Firm age and liquidity ratio as predictors of firm growth: Evidence from Indian firms. *Applied Economics Letters, 25*(19), 1373–1375.
Penrose, E. (2009). *The theory of the growth of the firm* (4th ed.). Oxford: Oxford University Press.
Ramaswamy, V., & Ozcan, K. (2016). Brand value co-creation in a digitalized world: An integrative framework and research implications. *International Journal of Research in Marketing, 33*(1), 93–106.
Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences, 2*(7), 1308.
Sampagnaro, G., & Lubrano Lavadera, G. (2013). *Identifying high-growth SMEs through balance sheet ratios*.
Segarra, A., & Teruel, M. (2014). The role of firm age and size in the growth of young firms: Evidence from Spain. *Research Policy, 43*(7), 1021–1038. https://doi.org/10.1016/j.respol.2014.03.005
Srhoj, S. (2022). Can we predict high growth firms with financial ratios? *Financial Internet Quarterly, 18*(1), 66–73.
Storey, D. J. (1994). *Understanding the small business sector*. London: Routledge.
Storey, D. J. (2011). Optimism and chance: The elephants in the entrepreneurship room. *International Small Business Journal, 29*(2), 303–322.
Van Witteloostuijn, A., & Kolkman, D. (2019). Is firm growth random? A machine learning perspective. *Journal of Business Venturing Insights, 11*, 1–5.
Wagner, J. (2003). Unobserved firm heterogeneity and the size-exports nexus: Evidence from German panel data. *Review of World Economics, 139*, 161–172.
Weinblat, J. (2018). Forecasting European high-growth firms- a random forest approach. *Journal of Industry, Competition and Trade, 18*(3), 253–294.