ЭКОНОМИКА
ECONOMY
ЭКОНОМИКА

PREDICTING HIGH-GROWTH FIRMS IN KAZAKHSTAN WITH MACHINE LEARNING METHODS

Yelzhas KADYR* *M.Sc in Economics., Senior Lecturer, International School of Economics, M. Narikbayev KAZGUU University, Nur-Sultan, Republic of Kazakhstan, e_kadyr@kazguu.kz*

Azat AITUAR *PhD in Economics, Assistant Professor, International School of Economics, M. Narikbayev KAZGUU University, Nur-Sultan, Republic of Kazakhstan, a.aituar@kazguu.kz, ORCID: 0000-0002-7625-8783, Scopus ID: 57280245800*

Saule KEMELBAYEVA *PhD in Economics, Associate Professor, International School of Economics, M. Narikbayev KAZGUU University, Nur-Sultan, Republic of Kazakhstan, s_kemelbayeva@kazguu.kz, ORCID: 0000-0002-7406-0589, Scopus ID: 57216337017*

DOI: 10.52123/1994-2370-2022-668
UDC 33:311
CICSTI 06.35.33

Abstract. In this paper, we study the effectiveness of popular machine learning methods for predicting high-growth firms in Kazakhstan and analyze this question with a set of 2012-2018 panel datasets. Moreover, we study the most important variables for the prediction of high-growth firms out of 50 variables included in the analysis. We develop a predictive design, where the past values are used to train classifiers that are applied in predicting future outcomes. Hereto, a test sample was used to evaluate the predictive performance of the classifiers. The results indicate that the best performing classifier increases the area under the curve equal to 0.8746. In terms of the variable importance, the firm's past growth in size, past growth in employment, past growth in revenue, and second derivative of the growth of financial variables contributed the most to predicting high-growth firms.
Keywords: high-growth firms, prediction, forecasting, machine learning, Kazakhstan.
JEL codes: C40, C60, C80, O40

Аңдатпа. Мақалада танымал машиналық оқыту әдістерінің Қазақстандағы жылдам дамып келе жатқан кәсіпорындарды болжау тиімділігі қарастырылады және 2012–2018 жылдарға арналған панельдік деректер жиынтығы талданады. Сонымен қатар, жылдам дамып келе жатқан кәсіпорындарды болжау үшін талдауға енгізілген 50 айнымалылардың ең маңыздылары қарастырылды. Нәтижелерді болжау үшін қолданылатын классификаторлар үйрететін және өткен мәндерді пайдаланатын болжамды дизайн әзірленді. Классификаторлардың болжамды тиімділігін бағалау үшін сынақ үлгісі пайдаланылды. Нәтижелер ең тиімді классификатор қисық астындағы ауданды 0,8746-ға ұлғайтатынын көрсетеді. Маңыздылығы тұрғысынан кәсіпорындардың бұрынғы өсуі, бұрынғы жұмыспен қамту өсімі, өткен табыстың өсуі және қаржылық айнымалылар өсуінің екінші туындысы жылдам дамып келе жатқан кәсіпорындарды болжауға себепші болды.
Түйін сөздер: жылдам дамып келе жатқан фирмалар, болжау, машиналық оқыту, Қазақстан.
JEL codes: C40, C60, C80, O40

* *Corresponding author: Y. Kadyr, e_kadyr@kazguu.kz*

Аннотация. В статье изучена эффективность популярных методов машинного обучения для прогнозирования быстрорастущих фирм в Казахстане, проанализирован набор панельных данных за 2012–2018 годы. Кроме того, изучены наиболее важные переменные для прогнозирования быстрорастущих фирм из 50 переменных, включенных в анализ. Разработан прогнозный дизайн, в котором прошлые значения используются для обучения классификаторов, которые применяются для прогнозирования будущих результатов. Для этого была использована тестовая выборка для оценки прогностической эффективности классификаторов. Результаты показывают, что наиболее эффективный классификатор повышает площадь под кривой, равную 0,8746. С точки зрения важности переменных, рост фирмы в прошлом, рост занятости в прошлом, рост выручки в прошлом и вторая производная от роста финансовых переменных в наибольшей степени способствовали прогнозированию быстрорастущих фирм.
Ключевые слова: быстрорастущие фирмы, предсказание, прогнозирование, машинное обучение, Казахстан.
JEL codes: C40, C60, C80, O40

1. Introduction

There is a substantial interest in predicting high-growth firms (HGFs) among policymakers, investors, and entrepreneurs *(Henrekson and Johansson, 2010; Mason and Brown, 2013; Goswami et al., 2019).* This interest arises because of HGFs' ability to create jobs, wealth, and their considerable contributions to productivity growth. Policymakers have an interest in employment, wages, and economic wealth, which are rising from growing business activity and entrepreneurship. A better understanding of predicting HGFs is important for investors who look to allocate funds to the right firms. Two crucial questions arise among policymakers: what kind of policies should be used to stimulate HGFs, and which firms should be focused on such policies. The nature of these questions is very different from each other. In particular, the first question on policy measures is about causality, as it assesses the measures to influence firm growth. The second question requires predicting firms that are most likely to show high growth at a later stage of their life cycle. Kleinberg et al. *(2015)* define this as a 'prediction policy problem'. Nevertheless, having the option to foresee HGFs does not imply that the growth of these firms can surely be influenced by some measures of policy. The predictive approaches give probabilities of outcomes, but they cannot answer the more complicated question of how to efficiently allocate resources, which is essential for policy decisions *(Athey, 2017).*

Current literature on HGFs focusing on regression models shows that it is difficult to accurately predict HGFs. Furthermore, there are significant number of research that do not use a proper predictive design. It has been stated that predicting potential HGFs is almost difficult due to the heterogeneous traits and stochastic nature of company growth, as demonstrated by Sterk et al. *(2021);* Coad et al. *(2014).* However, machine learning (ML) methods have shown effectiveness in predictive modeling for policy problems for difficult tasks and in diverse applications *(Athey, 2017; Kleinberg et al., 2015; Mullainathan and Spiess, 2017).*

This paper focuses on predicting HGFs in Kazakhstan and determinants of growth based on firm-level characteristics. To this end, we analyze commonly used ML techniques that can be applied to improve on HGFs predictions and compare them to standard regression methods. Besides, we study which predictors are the most important for predicting HGFs. Therefore, we focus on two research questions:

Which of the ML algorithm is the best in predicting the accuracy of HGFs?

Which predictors are the most important ones and how are they linked to the outcome variable?

The following is our research plan. A dataset of Kazakh firms is pre-processed and compiled. This firm-level dataset contains 50 variables based on literature about firm growth and is used to predict a high-growth firm's outcome. The number of employees and turnover are used to measure growth. We then use decision algorithms to train and tune the models. Finally, we evaluate the test sample's predictive performance. Our hypothesis is that machine learning approaches have a stronger predictive potential than classical regression methods.

This paper adds to the existing literature in several ways. First, Kazakhstan is an interesting country to study HGFs, because it has a transition economy and many firms are young. There is not enough research on HGFs in Kazakhstan. Second, we apply feature engineering techniques to increase the accuracy of the prediction. The weight of evidence (WOE), handling outliers,

scaling, and centering is examples of feature engineering techniques that were applied in this research. Third, feature selection was applied to increase accuracy and reduce computational cost. We use recursive feature elimination (RFE) as the wrapper feature selection method. Fourth, we add the second derivative of growth to the set of traditional features mentioned in the literature. This variable is one of the top important features in predicting HGFs for some classifiers. Next, we consider mAP (mean Average Precision), which is a popular evaluation metric used for object detection (classification). Finally, we were able to increase the area under the curve (AUC) to 87.46 % using the XGBoost method for auxiliary analysis, which is the highest value mentioned in the literature.

This paper is organized as follows. Section 2 provides a review of the literature related to high-growth firms. Section 3 describes the data and variables used in the analysis. Section 4 discusses the empirical framework. Section 5 presents the results of the predictions and Section 6 discusses the limitations of current research. Finally, a conclusion is provided in Section 7.

2. Methods and Data

We use the 2012–2018 yearly panel data set of firms in Kazakhstan by uniting two data sources: the Labor report by the Statistical Agency of Kazakhstan and the Report on the Financial and Economic Activities of the Enterprise, also by the Statistical Agency of Kazakhstan. A data-driven ML analysis should include as many relevant predictors as possible. Around 50 predictors were analyzed and we include two target variables for the baseline analysis. The main predictors are variables such as employment, the age of the firm, revenue, productivity, sales, business profit, etc. We perform the analysis independently for two different growth variables such that high employment growth and high revenue growth. The methodology for training and tuning the classifiers are described in this section. We look at a variety of ways for validating and evaluating their performance, as well as tools for determining variable importance. We use Hastie et al. *(2009)* technique for machine learning algorithms, which is sufficient for implementation purposes. The predictive model used in this study applies the validation set approach, which divides the entire data set into a training sample and a test sample for evaluating the prediction performance. The prediction findings achieved in the test sample are valid and reliable estimations of true out-of-sample performance, as illustrated in the next section of the study. To train classifiers, we used three machine learning algorithms: Lasso Regression, Random Forests, and XGBoost.

2.1 Related work

The literature on firm growth is a broad topic of interest in economics *(Coad and Tamvada, 2012; Davidsson and Delmar, 2006; Henrekson and Johansson, 2010; Storey, 1994; Delmar and Wiklund, 2008; Machado, 2016).* In this section, we only focus on the following relevant topics. First, we briefly review the empirical literature regarding factors affecting firm growth to find predictors of future HGFs. Second, to understand how my selection of variables is expected to perform on the data, we review the recent literature on the characteristics of HGFs. Lastly, we present a more detailed summary of several studies that have analyzed the identification of future HGFs.

2.2 Determinants of Firm Growth

The empirical research on factors that influence firm growth can be divided into two groups: internal and external factors *(Davidsson and Delmar, 2006).* Internal factors are divided into subcategories by Storey *(1994)* in terms of the entrepreneur, the firm, and the firm's strategy. Many variables under Storey's category of internal factors, including firm size, age, and legal form, have an impact on firm growth, according to empirical research. The majority of studies support similar conclusions for firm age and size, stressing the negative link between the two *(Davidsson and Delmar, 2006).* Moreover, Haltiwanger et al. *(2013)* show that when the firm age is controlled, there is no association between size and growth.

Entrepreneurial characteristics such as education, managerial experience, the number of founders, and functional competencies all have positive effects on firm growth *(Storey, 1994).* There is evidence of a link between entrepreneur goals and

visions and firm growth *(Delmar and Wiklund, 2008);* yet, most entrepreneurs have rather modest growth ambitions *(Gartner et al., 2004).*

In comparison to the two other internal components, the evidence in the domain of variables connected to corporate strategy is inconsistent. Storey's research, on the other hand, shows that market positioning, technological sophistication, and new product release all have a positive effect on firm growth and are employed more frequently than other variables. In-industry and regional factors are drivers of firm growth, according to studies using models based on external *factors (Capon et al., 1990; Delmar et al., 2003).* Furthermore, variables such as company-supporting policies in the form of innovation awards *(Wallsten, 2000),* networking and alliances *(Barringer et al., 2005),* internal and external funding opportunities (*Becchetti and Trovato, 2002; Beck and Demirguc-Kunt, 2006; Carpenter and Petersen, 2002),* and market and demand-related factors all have a favorable impact on firm growth *(Coad and Tamvada, 2012; Kangasharju, 2000).* On the effects of other external variables, there are no conclusive findings. Even while external characteristics play an important role in business growth, the majority of them are context-dependent and produce varied consequences depending on the configuration *(Davidsson and Delmar, 2006).*

Internal variables account for the majority of firm growth, although external factors also play a role. In the last decade, machine learning methods and more widely available processing capability have provided a solution to the problem of too many variables and interactions. As a result, in this study, we will use a vast number of potential predictor factors.

2.3 Observations about HGFs

The characteristics of HGFs have been studied intensively within the field of economics. Henrekson and Johansson *(2010)* reviewed most of these studies, showing mixed results. Seven HGF characteristics, on the other hand, have strong evidence *(Coad et al., 2014).* The first main character is the distribution of HGFs. The heavy-tailed distribution of firm growth has been studied by Bottazzi and Secchi *(2006)* and other authors. At the right end of the distribution, HGFs have received much interest. High-declining firms, on the other hand, have not attracted much interest. The second characteristic is that HGFs generate a large number of new employees. One of the motives for our paper is this characteristic. There is a lot of evidence for this result from many countries *(Acs and Mueller, 2008; Davidsson and Henrekson, 2002; Delmar et al., 2003)*. For example, when growth is calculated by employment using the OECD and EUROSTAT definitions for HGFs, HGFs represent 3-6 percent of total firms *(Hoffman and Junge, 2006).*

HGFs have a tendency to be youthful, but not necessarily small, according to the third *characteristic (Acs and Mueller, 2008; Daunfeldt et al., 2014).* HGFs are more widespread in high-tech industries, according to the fourth characteristic. The fifth characteristic is that high growth does not occur in a predictable pattern across time *(Delmar et al., 2003; Ho¨lzl and Janger, 2013).* When it comes to alternative growth computations, the sixth characteristic is that there is a distinct balance between defined HGFs *(Daunfeldt et al., 2014).* Finally, the seventh finding shows that prospective HGFs are challenging to *predict (Coad et al., 2014),* which is another important driving force behind our study.

2.4 Identifying Future High-Growth Firms

In most regression-based studies, predicting HGFs has challenges due to the difficulty of the task. ML approaches, on the other hand, have been shown to be beneficial in forecasting HGFs. We begin by looking at regression-based studies. By looking at balance sheet ratios as potential predictor variables for creating growth prediction models, Sampagnaro and Lubrano Lavadera *(2013)* contribute to the literature. They employ three regression models: Tobit with random and fixed effects, a combination of the two, and quantile regression. They discover that size, age, and internal cash flows are the most important determinants using Italian AIDA data from 21,182 enterprises between 2001 and 2008. However, there is no evaluation of forecast accuracy. Megaravalli and Sampagnaro *(2018)* apply a probit model to find the most significant factors for learning HGFs prediction models from balance sheet data.

They use a new data set of Italian firms (2010–2014), which only includes family-owned firms. They also employ an HGF definition that is based on two consecutive years of 20% annual revenue growth. According to the conclusions of this research, the most important financial variables are the liquidity ratio, solvency ratio, company age, cash flows, and working capital. With an AUC of 0.7078 on the ROC curve, the model's predictive performance is also addressed. The model's performance is determined entirely in-sample, as is typical of a variable important analysis. As a result, the reporting findings cannot be relied upon to estimate and compare model performance. Machine learning techniques have been used in several recent research to predict HGFs. Miyakawa et al. *(2017)* examined data from 1.7 million Japanese firms from 2006 to 2014 to forecast sales growth, profit growth, and firm's exit using a weighted random forest algorithm. They were able to attain an out-of-sample area under the ROC curve (AUC) of 0.68 and select 25 high-growth firms using a fixed probability threshold. Firm characteristics, region and industry, solvency score, and supply-chain network are all employed as predictors. The target variable is defined as a firm that achieves growth that is greater than the average growth of the predicted period by adding one standard deviation. The authors do not present any basis results that compare the model to standard baselines. Their research backs up the premise of using machine learning to forecast firm performance.

Weinblat *(2018)* forecasts European high-growth firms' performance using a random forest algorithm with 15 structural and financial indicators and defines the best applicable predictors in nine countries with 179,970 firms. The study presents the best out-of-sample prediction findings with an area under the ROC curve (AUC) of 0.8110 for Great Britain, using the (2004–2014) data set from the Amadeus database. If a firm's Birch-Schreyer employment growth indicator is in the top 10% of the sample, Weinblat *(2018)* uses a distributional high-growth characterization to evaluate if it belongs to the high-growth class. The Birch-Schrayer indicator takes into account the absolute and relative parts of employment growth. Weinblat *(2018)* does not find precise differences in predictability between groups of firms of different sizes in a complementary investigation. Moreover, the random forest algorithm presents a tool for evaluating variable importance, and the author's findings are consistent with previous research. Past growth, firm size, and age, based on Weinblat *(2018),* are the most important predictors across countries. According to Weinblat *(2018),* out-of-sample forecasts of HGFs are no longer beyond our capabilities, however the predictability of HGFs changes when a model is trained using firms from different nations. As a result, results cannot be applied to all of them. Furthermore, by considering features and algorithms, additional country-specific model enhancements can be produced.

By utilizing large data and computationally demanding methodologies, Coad and Srhoj *(2020)* contribute to the empirical study of high-growth enterprises. They measure growth using a binary indicator that identifies high-growth firms using the well-known Eurostat-OECD criteria *(Eurostat-OECD, 2007).* In their study, they look into how time-varying variables might be used to forecast high-growth. Kolkman and van Witteloostuijn *(2019)* use a dataset of 168,055 enterprises, only including basic demographic and financial information, to evaluate several machine learning algorithms to classic regression methods in terms of their goodness-of-fit. The random forest strategy achieves the best goodness-of-fit, while the innovative methods perform three to four times better overall. In addition, they develop four more proxies for personality and strategy factors based on 8,163 educational websites of Dutch SMEs. Our four text-analysis variables increase the R2 by around 2.5 percent. Bikowski and Antosiuk *(2021)* contrasted three algorithms: gradient boosting classifier, support vector machine, and logistic regression. Although they made the purposeful choice to use fewer predictors, they nevertheless managed to achieve highly promising precision, recall, and F1 scores for the best model, which were 57 percent, 34 percent, and 43 percent, respectively. The gradient boosting classifier produced the best results. By using contour plots to assess how size and age affect HGF chances, Coad and Karlsson *(2022)* add to the continuing discussion and offer a thorough empirical foundation for comprehending these

interactions in the big data era. Furthermore, we avoid a recurring issue in the literature where young and tiny enterprises are frequently underrepresented in commercial databases by using complete data on the firm population. They want to eventually offer a thorough field guide for other hunters on where to find gazelles as well as on the arid regions where they are unlikely to be discovered through this thorough analysis of the HGF distribution across firm size and age.

3. Identification of Future High-growth Firms

We use a 2012–2018 yearly panel data set of firms in Kazakhstan by uniting two data sources: the Labor report by the Statistical Agency of Kazakhstan and the Report on the Financial and Economic Activities of the Enterprise, also by the Statistical Agency of Kazakhstan. A data-driven ML analysis should include as many relevant predictors as possible. Around 50 predictors were analyzed and we include two target variables for the baseline analysis. The main predictors are variables such as employment, the age of the firm, revenue, productivity, sales, business profit, etc. We perform the analysis independently for two different growth variables such that high employment growth and high revenue growth.

To apply the ML analysis, we generated two data sets: a train data set and a test data set. Table 1 shows that each train data set consists of predictive variables from 2012 to 2014. Growth is observed between 2015 and 2017. The test data set consists of the exact predictive variables from 2013 to 2015 and growth is observed for the period between 2016 and 2018. The test data set is applied to measure the models' out-of-sample prediction performance. To improve data quality, all the incomplete cases are removed.

Table 1 - Train and test data sets

Year	2012	2013	2014	2015	2016	2017	2018
Train data	t-2 t-1 t Observation period			t+1 t+2 t+3 Growth period			
Test data		t-2 t-1 t Observation period			t+1 t+2 t+3 Growth period		

3.1 Identification of Future High-growth Firms

Information about a firms' growth state is crucial for developing prediction models and estimating its performance. HGFs are often measured using threshold measurements or relative measures Mason and Brown *(2010).* A firm is considered an HGF if its BirchSchreyer growth indicator values are in the top 10% of all firms over a three-year period. As a result, in terms of this index, this paper follows the technique of Schreyer *(2000),* Acs and Mueller *(2008),* and Lopez-Garcia and Puente *(2012).* They've all decided to take the top 10%. The following is how the index is calculated:

$$growth = (e_{t+3} - e_{t+1}) \cdot \frac{e_{t+3}}{e_{t+1}} \quad (1)$$

The number of employees at the firm in year t is e_t

$$Y_{ds} = \begin{pmatrix} 1, if\ growth_{ds} \geq q_{0.9,\,ds} \\ 0, if\ growth_{ds} < q_{0.9,\,ds} \end{pmatrix} \quad (2)$$

$q_{0.9,ds}$, ds is the 90th percentile of all growth values in the data set. HGFs are firms that have a growth value that is equal to or more than the 90 percent quantile. Low-

growth firms are referred to as ”others” (LGFs). The Birch-Schreyer growth indicator takes into account both absolute and relative employment growth.

According to Daunfeldt et al. *(2014),* finding HGFs in the manner described above results in rather unbalanced class distributions for Kazakh data, as seen in Table 2. The imbalance causes a number of methodological issues, which are discussed in Section 1. The proportion of HGFs in training and test samples is 10% in both baseline and auxiliary models. Table 2 shows these distributions together with the number of unique firms in training and test samples for various models.

Table 2 - The fraction of HGFs in training and test samples, and the number of unique firms in each analysis

Model	Training sample Firms	HGFs	Test sample Firms	HGFs
Employment	2273	235 (10.34%)	2241	231(10.31%)
Turnover	2303	248(10.77%)	2268	244(10.76%)

3.2 Predictors

We use feature selection by analyzing the literature for factors of firm growth based on Storey *(1994)* categorization of internal factors and Machado *(2016)* considering external factors affecting growth. As many relevant predictors as possible should be examined in a data-driven ML analysis. For the baseline analysis, we use a total of 50 predictors. Table 4 contains descriptive data as well as a summary table of target variables and predictors. Tables 5 and 6 show the frequency of Kazakhstani regions and industry types, respectively. In addition, the Appendix contains summary statistics for the generated variables.

The age of the firm is taken into account in this paper because, according to Harhoff et al. *(1998),* young firms expand quicker than older firms. Kumar and Ravi *(2007),* for example, depict the growth process of younger firms as having a significant variance. Almost all HGF studies take age and size into account. The firm’s age is calculated based on its initial appearance in the data. The firm’s size is measured by its total assets. Because larger firms are frequently older than smaller firms, there may be a link between these two factors. The number of tenges in a firm’s income is also used to determine its size, and productivity is computed by dividing revenue by employees. Another indicator of a company’s size is its workforce. Employment is used by Lopez-Garcia and Puente *(2012)* as a proxy for human capital, which is a key factor of HGFs. The Birch-Schreyer growth indicator is another key predictor. Following Lopez-Garcia and Puente *(2012),* we add previous growth to allow for auto-correlation *(2012).* Categorical characteristics such as the firm’s sector and region are taken into consideration. We include sixteen separate sectors based on NACE coding, as well as sixteen regions for branch and location control. There are other financial indicators to consider. Another factor to consider is the debit to-income ratio. It is a measurement of a company’s leverage, or how much of its overall funding is borrowed. Another metric that measures how effectively a firm uses its assets is a return on assets. It’s a metric that shows how profitable a firm is. The sales per employee ratio assess a firm’s employees’ ability to create sales and, as a result, their productivity *(Puri and Zarutskie, 2012).* The fixed assets ratio identifies the level of capital commitment. Schneider and Lindner *(2009)* believe that a high fixed-assets ratio can encourage a firm to increase in order to distribute its high fixed expenses across a greater number of goods.

Moreover, with the exception of age, we consider both absolute and relative changes for all continuous variables. The absolute changes are both regarded one year prior to the first year of growth Xt, with the first differences covering both one year $\Delta 1X = Xt - Xt-1$and two years $\Delta 2X = Xt - Xt-2$. The relative changes are both evaluated one year before the first year of the growth Xt, with the first differences covering both one and two years $l\Delta 1X = (Xt - Xt-1)/Xt-1$and two years $l\Delta 2X = (Xt - Xt-2)/Xt-2$. Instead of taking the first differences into account, we measured two

Birch-Schreyer growth indicators Δ1growth and $\Delta_2$growth a one-year and two-year time lag, respectively. There is no consensus in the literature about the time lag, as Coad and Tamvada *(2012)* pointed out. We also include second derivative growth features of the main variables. Given that we calculate features over three years, we incorporated a second derivative growth feature. It measures the increase or decrease of the relative growth in year one and year two versus the growth in year two and year three.

There are a lot of variables included in this analysis before pre-processing. The ML algorithms will choose the most important predictors and we will ignore other not relevant ones.

Table 3 - Listing and descriptions of the main variables

Variable	Description
Target variable	
High Growth (employment)	Binary: Leading 10%
High Growth (revenue)	Binary: Leading 10%
Predictors	
Age Employment Revenue Productivity Growth Debt ratio	Age of the firm (continuous) Number of personnel (continuous)
Sales per employee ratio Fixed assets ratio	Turnover in thousands tenge (continuous) Revenue/ Employment (continuous)
Size NACE	Birch-Schreyer growth indicator (continuous) Total debt/Total assets (continuous) Sales/Employment (continuous)
Region Legal form	Fixed assets/ Total assets (continuous) Total assets in thousands tenge (continuous) Categorical variable for different industries Categorical variable for regions Categorical variable for different legal forms

4. Empirical framework

The methodology for training and tuning the classifiers are described in this section. We look at a variety of ways for validating and evaluating their performance, as well as tools for determining variable importance. We use Hastie et al. *(2009),* Hsiao's *(2022)*, and Chen's *(2020)* techniques for machine learning algorithms for panel data, which are sufficient for implementation purposes. The predictive model used in this study applies the validation set approach, which divides the entire data set into a training sample and a test sample for evaluating the prediction performance. The prediction findings achieved in the test sample are valid and reliable estimations of true out-of-sample performance, as illustrated in the next section of the study. To train classifiers, we used three machine learning algorithms: Lasso Regression, Random Forests, and XGBoost[5].

Table 4 – Descriptive statistics before pre-processing of full data (N=24189)

Variable	Mean	Std. Dev	Min	Max
Target variables				
High growth (employment)	0.1015	0.3024	0	1
High growth (revenue)	0.1005	0.3007	0	1
Predictors				
Age	3.47	1.70	1	6
Employment	306.03	346.75	1	5236.5

[5] We did not use other machine learning algorithms such as Ridge regression, elastic net regression and neural network algorithms because prediction performance measurements were lower compared to Lasso Regression, Random Forests, and XGBoost.

Revenue	3536053	6134842	2413	9.07e+07
Productivity	12437.19	19115.49	20.81	215108.7
Growth (employment)	32.66	952.68	-847.67	81218.23
Growth (revenue)	2916539	1478162	-1.95e+07	5.99e+08
Debt ratio	0.2215	0.1897	0	0.8503
Sales per employee ratio	8900.7	12048.19	53.06	143842.6
Fixed assets ratio	0.3585	0.2473	0.0002	0.9253
Size	1.61e+07	3.15e+07	58371	3.98e+08
Return on sales	0.028	0.46	-12.84	0.9833

Table 5 – Frequencies of the regions of Kazakhstan

Region	Freq.	Percent
Akmola	1245	5.15
Aktobe	1100	4.55
Almaty city	5085	21.02
Almaty region	1178	4.87
Astana city	1998	8.26
Atyrau	1059	4.38
East Kazakhstan	1877	7.76
Karagandy	1967	8.13
Kostanay	1748	7.23
Kyzylorda	738	3.05
Mangystau	1045	4.32
North Kazakhstan	1235	5.11
Pavlodar	1041	4.30
South Kazakhstan	1378	5.70
West Kazakhstan	965	3.99
Zhambyl	530	2.19
Total	24189	100

Table 6 - Frequencies of the Industry type

NACE code	Freq.	Percent
Accomodation and food	606	2.51
Admin and support activities	1410	5.83
Agriculture	2773	11.46
Arts and entertainment	401	1.66
Construction	3334	13.78
Electricity and Gaz Supply	1191	4.92
Finance and Insurance	10	0.04
ICT	619	2.56
Manufacturing	5377	22.23
Mining	1063	4.39
Other services	87	0.36
Professional and scientific activities	1745	7.21
Real Estate	380	1.57
Transportation	1648	6.81

Water supply	691	2.86
Wholesale and Retail	2854	11.80
Total	24189	100

The prediction findings achieved in the test sample are valid and reliable estimations of true out-of-sample performance, as illustrated in the next section of the study. To train classifiers, we used three machine learning algorithms: Lasso Regression, Random Forests, and XGBoost.

4.1 Lasso Regression

Identifying future HGFs is considered a binary classification problem from the standpoint of statistical learning. Assume there is a vector of predictor variables x^i of firm i, where $x^i = x_{i1}, x_{i2}, ..., x_{ip}$ and a binary response outcome variable y_i (1 for HGFs, and 0 for non-HGFs). Given the variables of a certain firm, we must set the conditional probability p(y|x) of that firm referring to a class (0 or 1). Instead of explicitly computing the response y, the logistic regression uses a linear function of variables x to determine the likelihood that y belongs to a given class:

$$log\frac{\Pr(Y=1 \mid x)}{\Pr(Y=0 \mid x)} = \beta_0 + x^T\beta \quad (3)$$

where β_0 indicates the intercept, $\beta = (\beta_1, ..., \beta_p)$ indicates the linear coefficient, and Prob(Y = 1|x), Prob(Y = 0|x) denotes the conditional probabilities of the class labels 1 and 0, respectively. The log-likelihood can be written using the maximum likelihood method, which is frequently used in measuring coefficients:

$$\log(\beta_0,\beta) = \sum_{i=1}^{n} y_i \log \Pr(Y=1;\beta) + (1-y_i)\log(1-\Pr(Y=1;\beta)) = \sum_{i=1}^{n}\left(\beta_0 + X^T\beta\right) - \log\left(1+e^{\beta_0+X^T\beta}\right) \quad (4)$$

By placing an L1 constraint on β parameters, this logistic regression can be expanded into a Lasso-logistic regression *(Tibshirani, 1996; Friedman et al., 2001).* The challenge then becomes minimizing the negative log-likelihood function with the penalty term:

$$\sum_{i=1}^{n}\left(\log\left(1+e^{(\beta_0+X^T\beta)} - y_i\left(\beta_0 + X^T\beta\right)\right)\right) + \lambda\sum_{j=1}^{p} \mid \beta_j \mid \quad (5)$$

Because of the above constraint, making λ sufficiently large will result in the zeroing of multiple coefficients. A Lasso regression can have any number of variables depending on λ. As a result, both shrinkage and feature selection are produced at the same time. This characteristic also makes Lasso much easier to understand, making it a widely popular algorithm. The selection of relevant features is based on using statistical rather than theoretical reasons *(Coad and Srhoj, 2020).* The Lasso regression method is useful for dealing with big data sets and for filtering variables *(Friedman et al., 2010).*

4.2 Random Forests Algorithm

Breiman invented the Classification and Regression Tree (CART) in 1984, which was the predecessor of Random Forests (RF). In 1996, Breiman presented a new important approach for RF called Bagging Mills and Mills *(1990).* RF is a classification technique based on ensemble learning. It is built around two methods such that CART and Bagging. CART is a tree-structured classification method that makes decisions depending on a split of a variable in each node and works its way down until it reaches a leave node. The following is the CART growing algorithm. It iteratively splits each node into two sub-nodes by determining the best split variable and split value until the minimal node size is reached. CART has the advantage of being very well matched to data. When it comes to prediction, though, CART's accuracy isn't as high as it could be. To put it another way, CART has a low bias but a high variance *(Minsky and Papert, 1969).* To solve this difficulty, RF expands CART by providing the Bagging approach.

To begin with, this means that RF can fit a large number of CARTs into bootstrap sets re-sampled from the initial training set. Second, RF makes predictions based on the mode of the fitted CARTs' forecasts. Bagging will lower CART's variance while maintaining its low bias. Moreover, RF employs randomized node optimization to further decrease CART variance. All of RF's changes to CART have resulted in an excellent performance.

Aside from the low bias and low variance, RF has a number of other advantages. To begin, RF simply requires three parameters. They're relatively simple to tune if you stick to the recommended values. Second, RF produces an out-of-bag error, which is a good measure of the generalization error. Third, with training data, RF is resistant to irrelevant features and outliers *(Ho, 1998).*

Another advantage of RF is that it performs well when there is extremely skewed data, as Brown and Mues *(2012).* This is important for my research because HGFs account for only a low percentage of all firms. According to Yeh et al. *(2014),* the RF is less influenced by over-fitting than other ML methods. In addition, random sub-spacing significantly speeds up the calculating process Verikas et al. *(2015).* Unfortunately, due to the simultaneity of evaluating classification trees, RF is less transparent than one classification tree. Examining the RF's byproducts can help to counteract this. RF enables the evaluation of a global variable relevance ranking built on classification accuracy and unemployed observations of solitary trees *(Breiman, 2001; Verikas et al., 2015).*

4.3 Extreme Gradient Boosting (XGBoost)

Chen and Guestrin *(2016)* developed the eXtreme Gradient Boosting (XGBoost) technique, which has lately been popular in practical machine learning. It's a more efficient variant of gradient boosting decision trees, with the goal of improving speed and performance. Boosting is an ensemble algorithm that combines the outputs of "weak" learners to make them "strong." The goal of boosting is to sequentially train weak learners so that each consecutive tree tries to decrease the errors of the previous trees. Models are added in a logical order until no further improvements can be made. To create the final prediction, the predictions are pooled using a weighted average of regressions. Boosting is a non-parametric additive model that uses decision trees to build each additive element function. Extreme gradient boosting is a technique for creating a forest of trees in an additive way. The technique creates trees that minimize the prediction error iteratively. It generates the best set of predictive trees. If a new model is not suitable, new regression trees are combined progressively to reduce prediction error. The trees grow in a sequential manner, with each tree incorporating information from previous ones. To fit each tree, an adjusted version of the initial data is employed.

5. Results

The results are presented and analyzed in this section. First, we'll go over the outcomes of feature selection. After picking the most important features, the baseline model's outcomes are assessed and compared to existing literature using various performance indicators. We also give auxiliary prediction results for measuring firm turnover growth.

5.1 Feature selection for baseline analysis

We apply the wrapper method for feature selection. Figure 1 shows the results of the RFE method. The algorithm is set to investigate all possible subsets of the attributes. We can see the accuracy of the different attribute subset sizes in this plot. When all 50 attributes are considered the accuracy is at the highest level. Therefore, we consider all 50 attributes for RF and XGBoost. The Lasso regression is considered a separate feature selection method. Therefore, we consider only the top 29 important variables selected with the highest predictive power of HGF status including control variables such that age, region, and industry type (Figure 5).

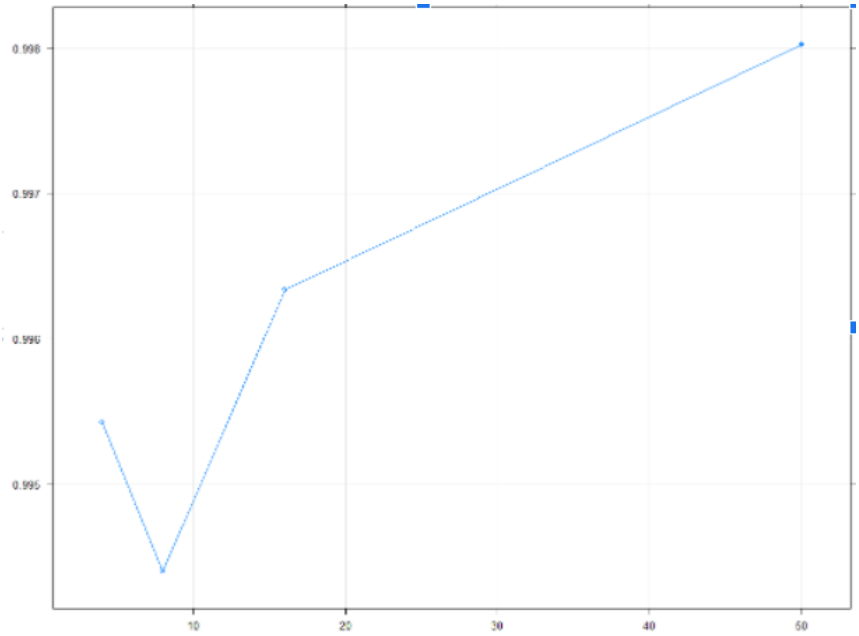

Figure 1 - Feature Selection Using Recursive Feature Elimination

5.2 Tuning hyper-parameter values of the Machine Learning Algorithms

The goal of the hyper-parameter tuning method is to discover the best set of parameters for the models. It is computationally expensive to search the entire space of possible values for parameters. As a result, we employed grid search and randomized search to identify and evaluate different sets of model parameters. In both of those strategies, we choose parameter values for experiments. For Lasso regression and RF of the restricted amount of parameters to tune, an intensive grid search was performed. Because of the large number of parameters, we used a randomized search to fine-tune the XGBoost model's performance. The performance of different model variations was tested using 10-fold cross-validation in both methods. The tuned hyper-parameter values for the three approaches are shown in Table 7.

The penalty parameter in Lasso regression is alpha. When applying an L1-norm constraint, some weights are set to zero in order to allow other coefficients to have nonzero values. Lasso regression can also be used to select features because the coefficients of less important features are decreased to zero. The first way to build a Lasso model is to determine the best lambda value. The alpha value for Lasso regression is one. 0.0272594 is the best cross-validated lambda in my analysis.

In the case of RF, the number of variables randomly chosen as candidates at each split (mtry) and the number of trees to develop is two tuning parameters to consider (ntree). There are plenty other parameters to consider. These two parameters, however, are the most likely to have the greatest impact on final accuracy. We used the default value of 500 for ntree. The number of variables randomly sampled as candidates at each split is equal to two, according to the grid search approach. This result is in line with Weinblat's *(2018)* findings for a number of countries, including Italy, Finland, the United Kingdom, and Poland. Moreover, Behr and Weinblat *(2017)* discovered that the maximum number of nodes per tree (maxnodes) is a helpful tuning parameter for preventing over-fitting. It stabilizes around 50 trees, according to the results of the out-of-bag (OOB) error (Figure 2). To reduce the computational burden, the number of trees is reduced to 50, as recommended by Breiman and Cutler *(2004).* Using 10-fold stratified cross-validations, the maximum node's value is discovered in a grid search (CVs). Because the grid search did not limit the number of maximum nodes, we set the maximum nodes to limitless. This study is in agreement with Weinblat's *(2018)* findings for a number of countries.

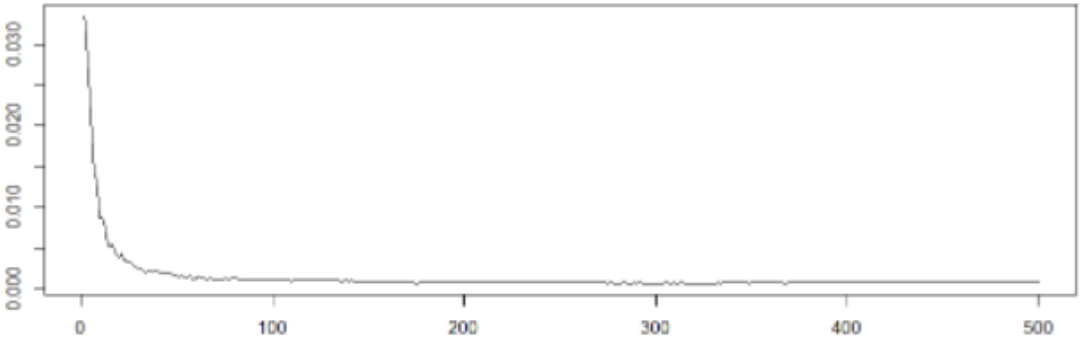

Figure 2 - OOB error

Table 7 also includes XGBoost's tuned parameters. We begin with the maximum tree depth allowed in each tree parameter. Max depth=3 has the best out-of-sample performance. The shrinkage parameter is set to 0.1, and the iterations are set to 150. These numbers cause lower errors and are in line with previous research.

Table 9 – Tuned hyperparameter values of the classifiers

Lasso Regression	XGBoost
alpha =1 lambda =0.0272594	Number of trees =150 Maximal tree depths =3 Shrinkage parameter =0.1 The minimum number of observations in trees' terminal nodes =10
Random Forests	
Number of randomly selected predictors =2 Number of trees =50 Minimal node size =unlimited	

5.3 Out-of-Sample Predictive Performance for baseline analysis

Table 10 shows the test sample's baseline results. The confusion matrices are provided in Appendix. RF and XGBoost outperform Lasso regression in practically every statistic. In terms of AUC, RF and XGBoost outperform Lasso regression classifiers by 0.0516 and 0.0839 points, respectively. This is a significant improvement. The RF and XGBoost algorithms, according to AUC's interpretation, rank a higher likelihood for a random Kazakh firm to be an HGF than a non-HGF, with probabilities of 79.79% and 83.02%, respectively.

Table 10 - Out-of-sample prediction results (10-fold cross-validation and SMOTE resampling)

Classifier	AUC	F-score	MAP	Accuracy	Sensitivity	FPR
Lasso Regression	0.7463	0.3123	0.2148	0.7511	0.5277	0.2237
XGBoost	0.8302	0.4611	0.3337	0.8587	0.5813	0.11
RF	0.7979	0.3711	0.3358	0.898	0.052	0.0067

The AUC of the RF classifier is similar to that of earlier literature; Weinblat *(2018)* produced an AUC of 0.81 for the United Kingdom, and Sharchilev et al. *(2018)* obtained an AUC of 0.85 for international data. It is, however, greater than the findings of Miyakawa et al. *(2017)* and Weinblat *(2018),* who found AUCs of 0.68 and 0.64 for Japanese and Finnish firms, respectively. Differences in these numbers can be argued in a variety of ways, depending on the methodology used and the variables used. There are no comparable studies on Kazakhstan.

Table 10 shows that the AUC difference between RF and XGBoost is 3.23 percent. The ROC curves of these classifiers are difficult to identify from each other (Figure 3). Furthermore, as compared to Lasso classifiers, the curves of RF and XGBoost

appear to be closer to the top left corner. The PR curves in Figure 4 demonstrate a comparable status based on performance exclusively in the positive class. The RF, XGBoost, and Lasso regression curves appear to be closer to the top right corner than the Lasso regression curve.

The AUC is the most widely used and reliable overall metric of predictive performance. Sensitivity and FPR are useful measurements for comparison and interpretation. With only a 0.67 percent (FPR) risk of misclassifying a non-high-growth firm as a high-growth firm, the RF classifier properly discovers 5.2 percent (sensitivity) of the high-growth firms. RF classifier has a very low sensitivity value. When compared to Weinblat's *(2018)* sensitivity and FPR values for European countries, the RF classifier exhibits lower sensitivity and FPR. In comparison to the values in Weinblat's *(2018)* study, XGBoost and Lasso regression show higher sensitivity and FPR.

The other supplied performance indicators in Table 10 are not comparable to past studies due to variations in class distributions. They're also reliant on the probability threshold that's used for prediction. For each classifier, the F-score optimization in training is used to identify it independently. On the other hand, these measurements inform decisions on the positive class that may be compared to one another. The F-score is particularly concerning, given the positive class is the more interesting. PPV and sensitivity should be evaluated jointly, as indicated in Figure 4, because their change is observed in a trade-off. Excessive values in one or the other are less important than keeping the two in balance. However, depending on the individual forecasting HGFs' goals and preferences, more weight can be given to accuracy in order to have more certainty in selecting a few more future HGFs or sensitivity in order to locate more future HGFs with lesser confidence.

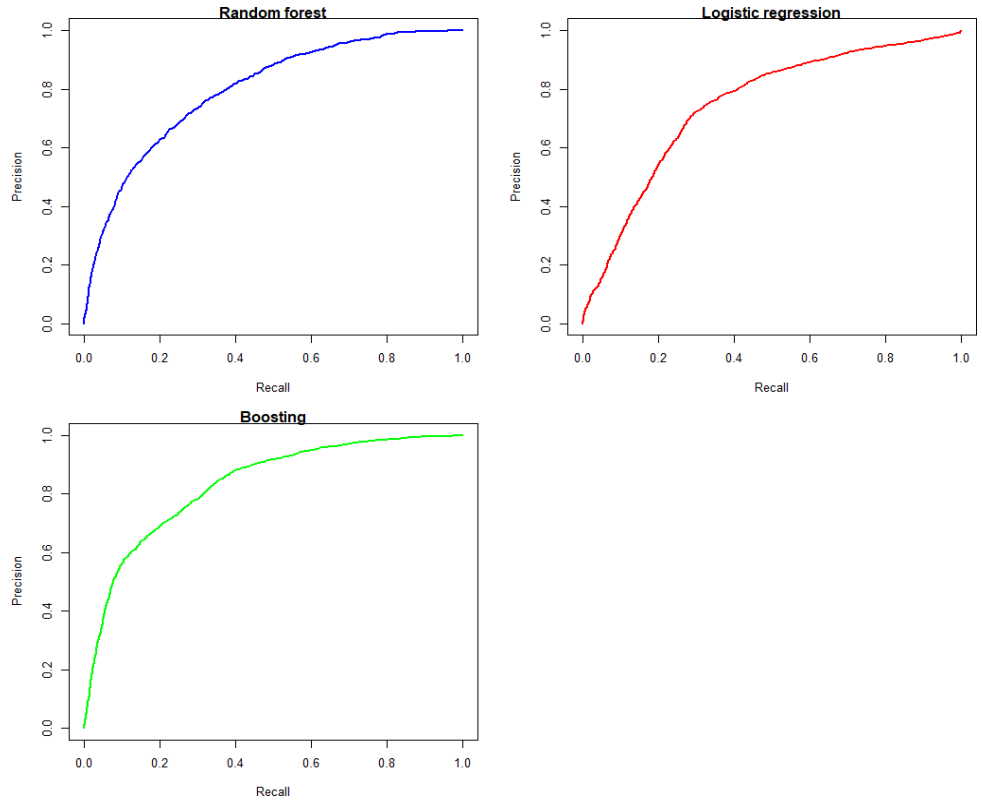

Figure 3 - ROC curves in the test sample

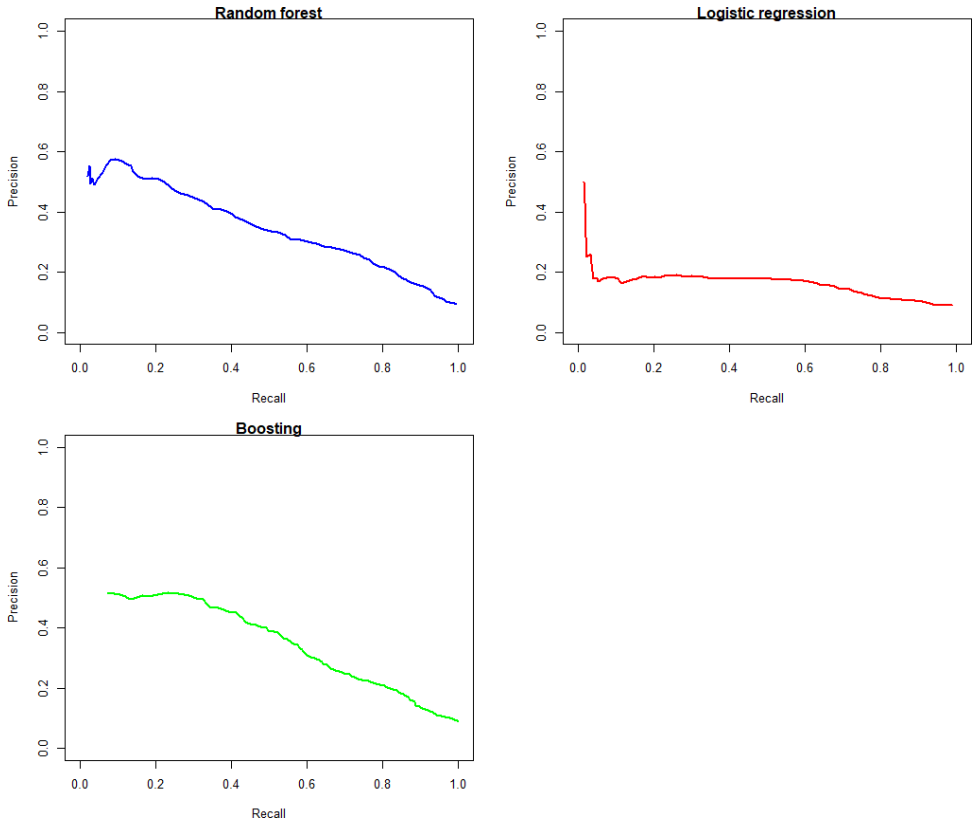

Figure 4 - PR curves in the test sample

The F-score ordering of the classifiers is the same as the AUC ordering. When compared to Lasso regression, RF and XGBoost perform better. Furthermore, by 0.09 points, XGboost has a higher F-score than RF. Because of the imbalanced class distribution, the overall accuracy measure in Table 8 is meaningless. Only by classifying all data as non-HGFs could a classifier achieve over 95 % accuracy. This metric is included, however, because of its widespread use in the literature. It's difficult to pick just one strategy for forecasting HGFs because different measurements of predictive performance yield varied results. Based on the results, we can see that RF and XGBoost outperform Lasso regression in terms of predictive performance. When comparing RF with XGBoost, the latter outperforms the former in every category except accuracy and MAP. In the end, it all comes down to a preference for the measure, which might originate from policymakers' and investors' needs.

5.4 Evidence on the Most Meaningful Predictors

Figures 5, 6, and 7 show the findings of variable importance for classifiers. The horizontal axis depicts the relative importance. On the vertical axis, the most essential variables are listed in descending order from the top. In Lasso regression, the four most important variables are past growth in employment (growthdelta1), fixed assets ratio (far), revenue, and size. In XGBoost, the most important predictors are past growth in employment (growthdelta1), employment (emp), age, and second derivatives of growth in employment (growth2delta). The three most important variables after these four are the past two years of growth in employment, past relative growth in return on sales (rosldelta1), and past growth in revenue (revenuedelta1). The top seven predictors for RF are past growth in employment (growthdelta1), the second derivative of growth in employment (growth2delta), past two years of growth in employment (growthdelta2), employment, past growth in revenue (revenuedelta1), size, and past two years growth in size (sizedelta2). The predictors such that past growth in employment and revenue are located in the top 10 of each classifier. The variables of past growth in employment and employment are important for RF and XGBoost. These results are consistent with Weinblat's *(2018)* findings. They are also consistent with past research.

For some tree-based ML approaches, partial dependence plots (PDP) present valuable interpretations for variable importance. In Figure 8, we use the RF classifiers to plot the 9 most important predictors. Other classifiers show the same patterns. The horizontal axis depicts the predictors' centered and scaled values, while the vertical axis depicts the probability of classifying a firm as an HGF given all other features.

If there is more variance in the plot for any specific predictor variable, it indicates that the value of that variable has a significant impact on the model, however, if the line is constant near zero, it indicates that the variable has no impact on the model. Single variables demonstrate how their values affect the model; on the y-axis, a negative value for a predictor variable indicates that it is less likely to predict the proper class on that observation, while a positive value indicates that it has a positive impact on predicting the correct class. Past growth indicators such as past growth in employment, past growth in revenue, and past growth in size appear to contribute the most and demonstrate the highest conditional variation in partial dependency when combined with variable importance analysis.

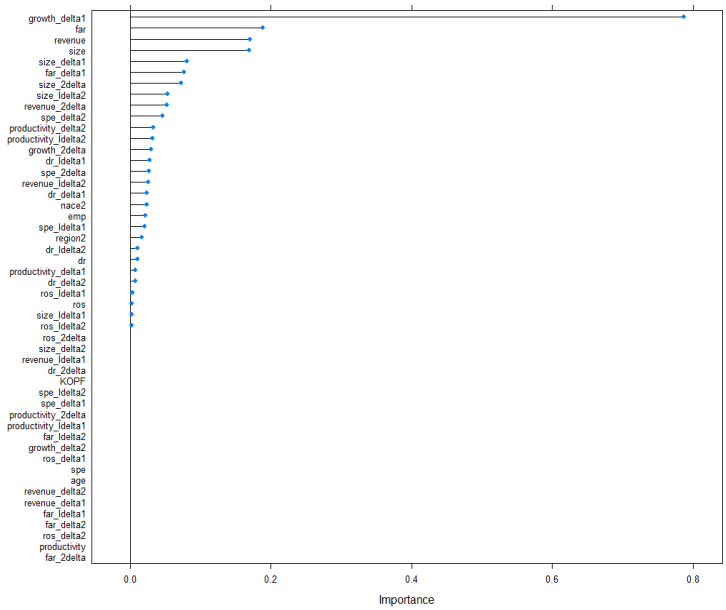

Figure 5 - Variable Importance (Lasso Regression)

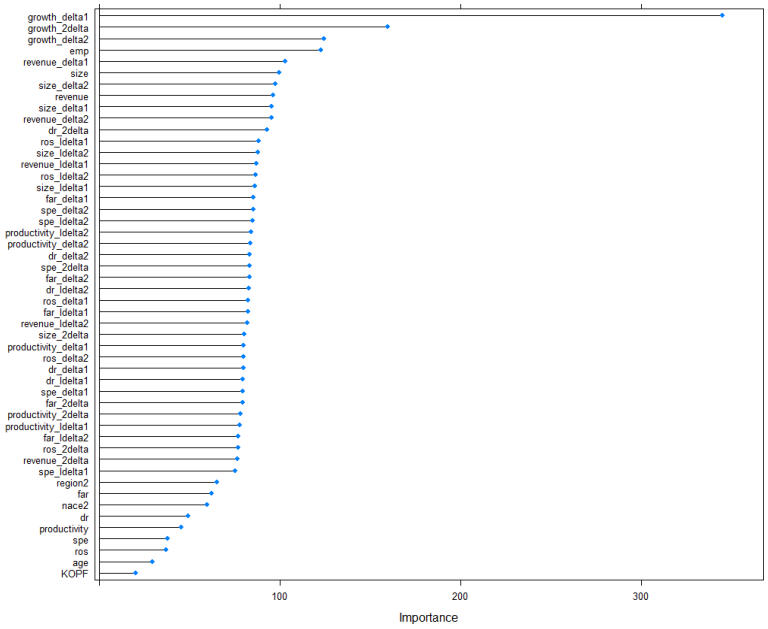

Figure 6 - Variable Importance (RF)

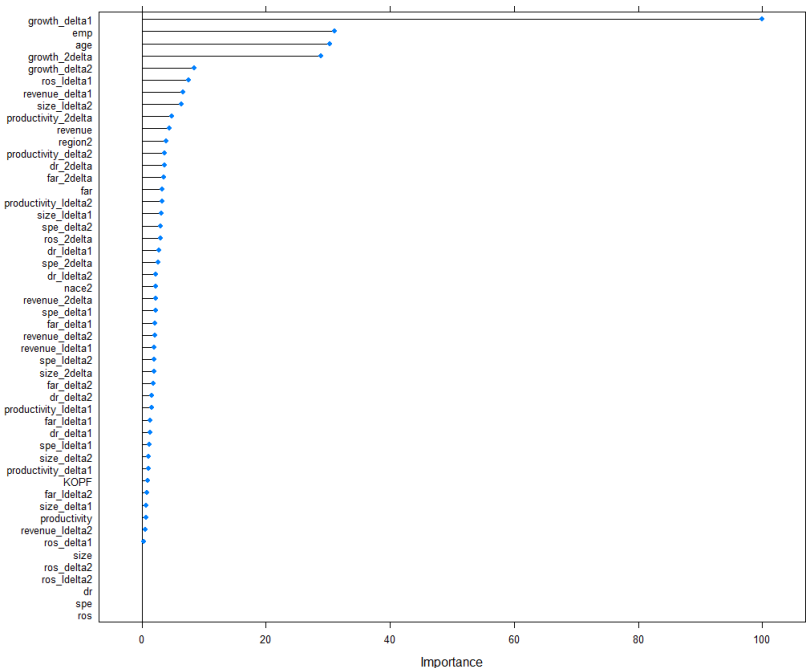

Figure 7 - Variable Importance (XGBoost)

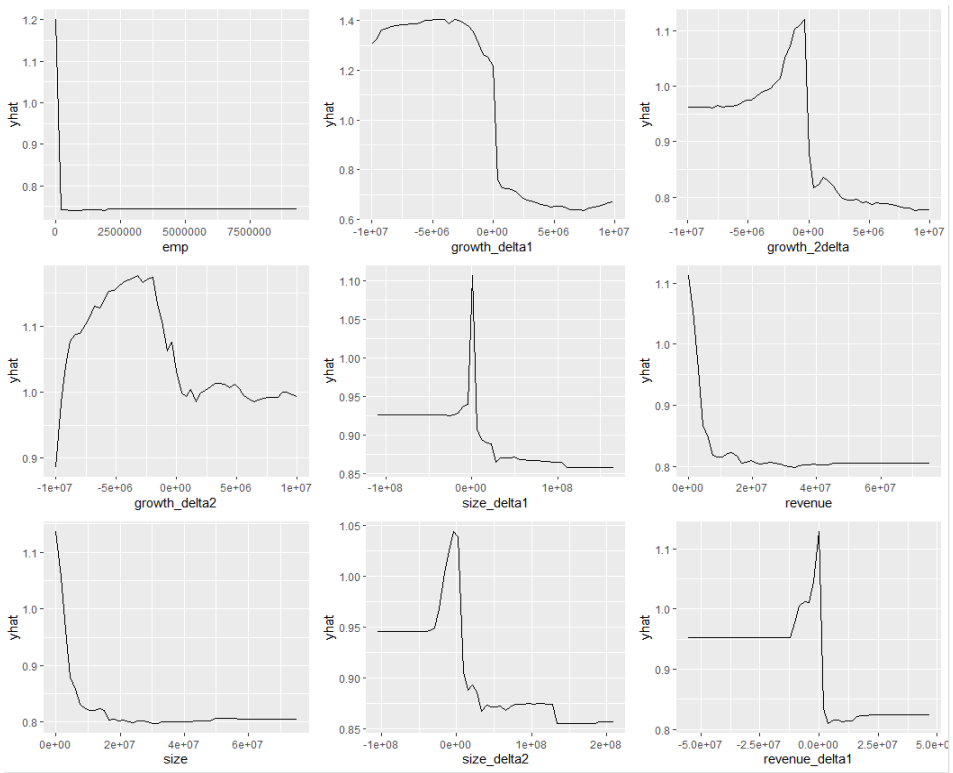

Figure 8 - Partial Dependence Plots (Random Forest)

6. Discussion

In this section we discuss the implications of our findings as well as the study's limitations. Our findings raise a number of issues that have policy consequences. In discovering HGFs, the baseline study demonstrates that RF and XGBoost outperform Lasso regression in terms of predictive performance. The classifiers perform in a variety of ways, but they all detect a modest number of HGFs. They do, however, have a high level of precision. In general, our findings support prior findings that ML approaches may be used to solve prediction policy problems in the same way they have been used in other areas *(Mullainathan and Spiess, 2017).*

When compared to traditional methodologies, the improvements in performance with ML algorithms are significant. However, we can observe that for a prediction problem, ML algorithms are unable to find superior relationships from data *(Coad et al., 2014).* Moreover, the findings on the most important predictors are in line with earlier research *(Weinblat, 2018).* We can see that firm revenue, size, past change, and firm age are the most important predictors. The auxiliary analysis reveals several interesting facts that policymakers and investors should be aware of. First, it appears that prediction accuracy is vulnerable to the high-growth definition, based on the results of a robustness check. When measuring revenue growth rather than employment growth, this prediction problem is easier to address. There are less features used in the auxiliary analysis compared to the baseline model. It is vital to notice the discrepancies in growth measurements, HGF definitions, and HGF distributions, according to earlier work *(Daunfeldt et al., 2014; Delmar et al., 2003).* Previous research has used a variety of HGF definitions to calculate the number of HGFs in a given economy *(Henrekson and Johansson, 2010).* In general, three types of inputs (investment, employees), values (assets, market capitalization), and outputs (sales turnover, profits) can be used to measure firm growth *(Garnsey et al., 2006).* Due to the minimal overlap between these measures of growth, a firm may be classed as having high growth in terms of sales turnover but low growth in terms of employment, or vice versa *(Delmar et al., 2003).* We choose to use both number of employees and turnover as growth indicators when studying growth persistence because we recognize that a single metric will not capture all aspects of firm growth *(Janssen, 2009).* However, it is more interesting to use turnover since it shows the effect of expansion on a firm. Furthermore, increasing employment is rarely, if ever, a firm owner's goal, although increasing sales and turnover is *(Dobbs and Hamilton, 2007).* There are some questions that arise based on the reasoning above. Why is it so difficult to forecast HGFs, and are there any possible solutions? While the first is outside the focus of this paper, the identical issue has been raised in the literature before *(Coad et al., 2014).* The reasons for the difficulties are mostly due to the heterogeneity of firms and how they grow. Furthermore, there are a number of characteristics that have been linked to firm growth but for which there is no high-dimensional data. Additional methodological improvements can be developed based on the technique taken in this paper to potentially improve predictive accuracy. Increased data quantity and quality, on the other hand, are likely to yield the most promising results.

The previous paragraph's ideas and subjects are considered part of a basic guideline for future work on identifying HGFs. However, there are a few limitations to the analysis presented in this paper, which can be transformed into particular research questions for the future. ML techniques require a set of considerations about the training process from a methodological standpoint. If time were not an issue, we could enhance prediction performance by trying different re-sampling schemes and tuning parameter values. However, tuning usually only yields minor gains. Increasing the number and quality of data and variables could be a more effective strategy. We could add some additional features about CEOs of the firms such that gender, education, managerial experience, and the number of founders of the firms, which are proven to have a substantial impact on a firm's growth *(Guzman and Stern, 2015).* The data sample employed in this paper is considerable, with 50 predictors, yet it is little in comparison to what machine learning algorithms can handle.

In terms of detecting HGFs, we think the last argument is the most promising. Given ML approaches' capacity to handle non-traditional data forms like text *(Kolkman and van Witteloostuijn, 2019),* we believe news stories could be useful in HGF prediction, for example. Sharchilev et al. *(2018)* used a similar strategy with promising results. More research towards improving the prediction performance of future HGFs is needed in the long run. If a reasonable level of prediction accuracy is achieved, the question of how to distribute resources best and with what instruments must be addressed to the same degree. Furthermore, rather than pure prediction, this requires interference investigations *(Athey, 2017).*

7. Conclusion

We used a predictive approach in this paper that is analogous to a real forecasting scenario, in which previous values of key variables predict uncertain future events. We trained three classifiers to predict HGFs in a 2012–2017 learning sample of Kazakh firms using complex ML algorithms and a wide number of predictors. These classifiers' predictive performance was evaluated in out-of-sample test years from 2013 to 2018. According to the findings, RF and XGBoost outperform Lasso regression in predicting HGFs in the baseline model. XGBoost was shown to be the best performing classifier, with an out-of-sample improvement of 0.084 points above the Lasso regression in terms of AUC. When revenue growth is taken into account, however, there appears to be no significant improvement in XGBoost and RF compared to the Lasso regression. However, XGBoost is the best performing classifier with AUC equal to 0.8746. This responds to Section 1's first research question. In terms of important variables (question 2), the firm's past size growth, past employment growth, and past growth in revenue have the biggest effect on forecasting HGFs in the baseline model. Furthermore, the findings revealed that identifying HGFs was easier and computationally effective when turnover was measured rather than employment. By applying feature selection, we consider only 16 important variables including the second derivative of growth the feature of main financial variables. Further research is needed to evaluate the importance of the second derivative of growth features in predicting HGFs.

This paper's empirical framework has several limits in terms of data amount and quality, as well as room for development in methodological choices. Where more research to improve the predictive scheme used in this paper is needed, causal studies are needed to answer the question of how to best deploy resources for prospective HGFs and with what instruments. Nonetheless, we present a robust machine learning-based prediction scheme with policy-relevant outcomes. With the data given, we find that the ML approaches are effective and have significant improvement compared to traditional econometric methods.

REFERENCES

Acs, Z. J., Mueller, P. (2008). Employment effects of business dynamics: Mice, gazelles, and elephants. *Small Business Economics.* 30 (1), 85–100.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. Science 355 (6324), 483–485.

Barringer, B.R., Jones, F.F., Neubaum, D.O. (2005). A quantitative content analysis of the characteristics of rapid-growth firms and their founders. *Journal of business venturing*. 20 (5), 663–687.

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.

Beck, T., Demirguc-Kunt, A. (2006). Small and medium-sized enterprises: Access to finance as a growth constraint. *Journal of Banking & Finance.* 30 (11), 2931–2943.

Behr, A., Weinblat, J. (2017). Default patterns in seven EU countries: A random forest approach. *International Journal of the Economics of Business.* 24 (2), 181–222.

Breiman, L. (2001). Random forests. *Machine learning.* 45 (1), 5–32.

Breiman, L., Cutler, A. (2004). Random forest-manual. http://www.stat.Berkeley.edu/˜Breiman/RandomForests/ccmanual.htm.

Brown, I., Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications.* 39 (3), 3446– 3453.

Capon, N., Farley, J.U., Hoenig, S. (1990). Determinants of financial performance: a meta-analysis. *Management Science.* 36 (10), 1143–1159.

Carpenter, R. E. and B. C. Petersen (2002). Is the growth of small firms constrained by internal finance? *Review of Economics and Statistics.* 84 (2), 298–309.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chen, J. (2020, October). Workshop 2-introductory machine learning: decision trees, random forests, and other dendrological methods. *In 90th International Atlantic Economic VIRTUAL Conference. IAES*.

Coad, A., Daunfeldt, S.-O., H¨olzl, W., Johansson, D., Nightingale, P. (2014). High-growth firms: introduction to the special section. *Industrial and Corporate Change.* 23 (1), 91–112.

Coad, A., Tamvada, J.P. (2012). Firm growth and barriers to growth among small firms in India. *Small*

Business Economics. 39 (2), 383–400.
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.
Coad, A., Karlsson, J. (2022). A field guide for gazelle hunters: Small, old firms are unlikely to become high-growth firms. *Journal of Business Venturing Insights*, 17, e00286.
Daunfeldt, S.-O., N. Elert, and D. Johansson (2014). The economic contribution of high growth firms: do policy implications depend on the choice of growth indicator? *Journal of Industry, Competition and Trade*. 14 (3), 337–365.
Davidsson, P., Delmar, F. (2006). High-growth firms and their contribution to employment: The case of sweden 1987–96. *Entrepreneurship and the growth of firms*, 156–178.
Davidsson, P., Henrekson, M. (2002). Determinants of the prevalence of start-ups and high-growth firms. *Small business economics.* 19 (2), 81–104.
Delmar, F., Davidsson, P., Gartner, W.B. (2003). Arriving at the high-growth firm. *Journal of business venturing.* 18 (2), 189–216.
Delmar, F., Wiklund, J. (2008). The effect of small business managers' growth motivation on firm growth: A longitudinal study. *Entrepreneurship theory and practice.* 32 (3), 437–457.
Dobbs, M., Hamilton, R.T. (2007). Small business growth: recent evidence and new directions. *International journal of entrepreneurial behavior & research.*
Fawcett, J. (2005). Criteria for evaluation of theory. *Nursing science quarterly.* 18 (2), 131–135.
Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software.* 33 (1), 1.
Friedman, J., T. Hastie, R. Tibshirani, et al. (2001). The elements of statistical learning, *Springer series in statistics New York*. 1.
Garla, V., Taylor, C., Brandt, C. (2013). Semi-supervised clinical text classification with Laplacian svms: an application to cancer case management. *Journal of biomedical informatics.* 46 (5), 869–875.
Garnsey, E., Stam, E., Heffernan, P. (2006). New firm growth: Exploring processes and paths. *Industry and Innovation.* 13 (1), 1–20.
Gartner, W.B., Gartner, W.C., Shaver, K.G., Carter, N.M., Reynolds, P.D. (2004). *Handbook of entrepreneurial dynamics: The process of business creation.* Sage.
Goswami, A.G., Medvedev, D., Olafsen, E. (2019). *High-growth firms: Facts, fiction, and policy options for emerging economies.* World Bank Publications.
Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*. 3 (Mar), 1157–1182.
Guzman, J., Stern, S. (2015). Where is silicon valley? *Science.* 347 (6222), 606–609.
Haltiwanger, J., Jarmin, R.S., Miranda, J. (2013). Who creates jobs? small versus large versus young. *Review of Economics and Statistics.* 95 (2), 347–361.
Harhoff, D., Stahl, K., Woywode, M. (1998). Legal form, growth and exit of west German firms—empirical results for manufacturing, construction, trade and service industries. *The Journal of industrial economics.* 46 (4), 453–488.
Hastie, T., Tibshirani, R., Friedman, J. (2009). Unsupervised learning. *In The elements of statistical learning*, pp. 485–585. Springer.
Henrekson, M., Johansson, D. (2010). Gazelles as job creators: a survey and interpretation of the evidence. *Small business economics,* 35 (2), 227–244.
Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence,* 20 (8), 832–844.
Hoffman, A., Junge, M. (2006). Documenting data on high-growth firms and entrepreneurs across 17 countries. fora. *Technical report*, Copenhagen: Mimeo.
Ho¨lzl, W., Janger, J. (2013). Does the analysis of innovation barriers perceived by high-growth firms provide information on innovation policy priorities? *Technological Forecasting and Social Change*, 80 (8), 1450–1468.
Hsiao, C. (2022). *Analysis of panel data.* Cambridge university press.
Janssen, F. (2009). The conceptualization of growth: are employment and turnover interchangeable criteria? *The Journal of Entrepreneurship,* 18 (1), 21–45.
Kangasharju, A. (2000). Growth of the smallest: Determinants of small firm growth during strong macroeconomic fluctuations. *International Small Business Journal,* 19 (1), 28–43.
Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105 (5), 491–95.
Kolkman, D., van Witteloostuijn, A. (2019). *Data science in strategy: Machine learning and text analysis in the study of firm growth.*
Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, Springer.
Kumar, P.R., Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–a review. *European journal of operational research.* 180 (1), 1–28.

Lipton, Z.C., Elkan, C., Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize f1 measure. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 225–239. Springer.
Lopez-Garcia, P., Puente, S. (2012). What makes a high-growth firm? a dynamic probit analysis using Spanish firm-level data. *Small Business Economics.* 39 (4), 1029–1041.
Machado, H.P. (2016). Growth of small businesses: a literature review and perspectives of studies. *Gest˜ao & Produ¸c˜ao.* 23, 419–432.
Mason, C., Brown, R. (2010). *High growth firms in Scotland*.
Mason, C., Brown, R. (2013). Creating a good public policy to support high-growth firms. *Small business economics,* 40 (2), 211–225.
Megaravalli, A. V. and G. Sampagnaro (2018). Firm age and liquidity ratio as predictors of firm growth: evidence from Indian firms. *Applied Economics Letters,* 25 (19), 1373– 1375.
Mills, T.C., Mills, T.C. (1990). *Time series techniques for economists*. Cambridge University Press.
Minsky, M., Papert, S. (1969). *An introduction to computational geometry*. Cambridge tiass. HIT
Miyakawa, D., Miyauchi, Y., Perez, C. (2017). *Forecasting firm performance with machine learning: Evidence from Japanese firm-level data*. RIETI.
Mullainathan, S., Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives.* 31 (2), 87–106.
Puri, M., Zarutskie, R. (2012). On the life cycle dynamics of venture-capital-and non-venture-capital-financed firms. *The Journal of Finance.* 67 (6), 2247–2293.
Sampagnaro, G., Lubrano Lavadera, G. (2013). *Identifying high-growth SMEs through balance sheet ratios*.
Schneider, O., Lindner, A. (2009). The value of lead logistics services. *In IFIP International Conference on Advances in Production Management Systems*, pp. 315–322. Springer.
Schreyer, P. (2000). *High-growth firms and employment.*
Sharchilev, B., Ustinovskiy, Y., Serdyukov, P., Rijke, M. (2018). Finding influential training samples for gradient boosted decision trees. *In International Conference on Machine Learning*, pp. 4577–4585. PMLR.
Sterk, V., Sedl´a˘cek, P., Pugsley, B. (2021). The nature of firm growth. *American Economic Review.* 111 (2), 547–79.
Storey, D. (1994). Understanding the small business sector. *The University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.*
Sun, Y., Wong, A.K., Kamel, M.S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence.* 23 (04), 687–719.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 58 (1), 267–288.
V. Uloza, Padervinskis, E. (2015). Data-dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: acoustic versus contact microphone. *Medical engineering & physics.* 37 (2), 210–218.
Wallsten, S. J. (2000). The effects of government-industry R&D programs on private R&D: the case of the small business innovation research program. *The RAND Journal of Economics*, 82–100.
Weinblat, J. (2018). Forecasting European high-growth firms-a random forest approach. *Journal of Industry, Competition and Trade.* 18 (3), 253–294.
Yeh, C.-C., Chi, D.-J., Lin, Y.-R. (2014). Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences.* 254, 98–110.
Żbikowski, K., Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, 58(4), 102555.

МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІМЕН ҚАЗАҚСТАНДАҒЫ ЖОҒАРЫ ӨСЕТІН ФИРМАЛАРДЫ БОЛЖАУ

Елжас ҚАДЫР*, экономика ғылымдарының магистры, Халықаралық экономика мектебінің аға оқытушысы, М. Нәрікбаев атындағы КАЗГЮУ университеті, Нұр-Сұлтан, Қазақстан Республикасы, e_kadyr@kazguu.kz*
Азат АЙТУАР*, PhD (экономика), ассистент-профессор, Халықаралық экономика мектебі, М. Нәрікбаев атындағы КАЗГЮУ Университеті, Нұр-Сұлтан, Қазақстан Республикасы, a.aituar@kazguu.kz, ORCID: 0000-0002-7625-8783, Scopus ID: 57280245800*
Сауле КЕМЕЛЬБАЕВА*, PhD (экономика), қауымдастырылған профессор, Халықаралық экономика мектебі, М. Нәрікбаев атындағы КАЗГЮУ Университеті, Нұр-Сұлтан, Қазақстан Республикасы, s_kemelbayeva@kazguu.kz, ORCID: 0000-0002-7406-0589, Scopus ID: 57216337017*

ПРОГНОЗИРОВАНИЕ БЫСТРОРАСТУЩИХ ФИРМ В КАЗАХСТАНЕ С ПОМОЩЬЮ МЕТОДОВ

МАШИННОГО ОБУЧЕНИЯ

Елжас КАДЫР, *магистр экономических наук, старший преподаватель, Международная школа экономики, Университет КАЗГЮУ им. М. Нарикбаева, Нур-Султан, Республика Казахстан, e_kadyr@kazguu.kz*

Азат АЙТУАР, *PhD по экономике, ассистент-профессор, Международная школа экономики, Университет КАЗГЮУ им. М. Нарикбаева, Нур-Султан, Республика Казахстан, a.aituar@kazguu.kz, ORCID: 0000-0002-7625 -8783, ScopusID: 57280245800;*

Сауле КЕМЕЛЬБАЕВА, *PhD по экономике, ассоциированный профессор, Международная школа экономики, Университет КАЗГЮУ им. М. Нарикбаева, Нур-Султан, Республика Казахстан, s_kemelbayeva@kazguu.kz, ORCID: 0000-0002-7406-0589, Scopus ID: 57216337017*