Enhancing Effectiveness of Government Credit Programs for SMEs: A Machine Learning Approach*

Minho Kim[†] Korea Development Institute

Youngdeok Hwang[‡] Baruch College, City University of New York

Abstract

Government-subsidized loans and public credit guarantees are key policies that provide support to Small and Medium-sized Enterprises experiencing financial difficulties, particularly those with high growth potential but facing tight credit constraints. However, the effectiveness of these credit programs can be compromised by adverse selection and poor screening of beneficiary firms. To tackle this challenge, we propose a data-driven approach that aims to allocate credits to more viable and credit-constrained firms. We employ machine learning techniques, utilizing a range of firm-specific characteristics, to predict the annual sales growth rate. Our approach substantially improves government credit programs by effectively identifying suitable target firms, and can be easily applied to other business support programs. Additionally, our findings indicate that firm size and age serve as highly informative indicators of credit constraints.

JEL Codes: G21, G28, H81, L26

Keywords: Subsidies, Loan guarantees, Machine learning, Small and medium enterprises, Financing constraints, Program evaluation.

*Most of this work was done while Kim was at Baruch College as a visiting scholar.

[†]Email: minhokim@kdi.re.kr

[‡]Email: youngdeok.hwang@baruch.cuny.edu

1 Introduction

Government-subsidized loans and public credit guarantee programs are important and common industrial policies in many countries to support Small and Medium-sized Enterprises (SMEs) facing financial challenges. These programs facilitate the growth of SMEs by providing them access to capital. SMEs, especially startups and small enterprises, often struggle to secure funding due to their limited ability to demonstrate their creditworthiness, resulting in high interest rates or loan denials. Government financing addresses credit-market failures caused by information asymmetry, by aiding credit-constrained SMEs in expanding their businesses.

However, despite policymakers' intention to support firms experiencing temporary credit constraints but with high growth potential, these programs may inadvertently attract a disproportionate number of "lemon" beneficiaries who do not genuinely need the support. Both financially constrained and unconstrained firms have an incentive to take advantage of subsidized loans due to their lower interest rates compared to private alternatives. Moreover, banks may engage in adverse selection by favoring riskier borrowers, and their screening and monitoring of beneficiary firms may be less rigorous (moral hazard, Cowan et al., 2015; De Blasio et al., 2018; Lagazio et al., 2021) because public credit programs guarantee most of the debts (e.g., 85 - 100% in South Korea).

These challenges from both borrowers and lenders often lead to a sub-optimal selection of beneficiary firms, which adds a burden to public finance without achieving the intended outcomes. The fundamental challenge is clear: How can we identify the firms that are currently experiencing credit-constraints and yet would be greatly propelled by some government support?

This paper demonstrates that government credit programs can be significantly improved through the adoption of a data-driven approach, specifically by utilizing a machine learning

model to predict annual sales growth rates. By leveraging a comprehensive dataset related to policies targeting young firms, we show that allocating credits to firms with higher growth potential and greater credit constraints can be a straightforward yet efficient way to weed out the poor candidates. We combine four large-scale administrative databases covering nationwide government-guaranteed and direct government loans provided between 2010 and 2015 with individual firm-level annual financial and basic characteristics data. This integration allows us to create a longitudinal dataset at the firm level. The individual firm information includes the financial balance-sheet information from 2009 to 2017 and other relevant characteristics, sourced from Korea Enterprise Data (KED), the largest database on Korean SMEs.

Our research holds significant relevance to current policy shift, as many governments are transitioning toward data-driven decision-making in their public services and calling for incorporating Artificial Intelligence (AI, Berryhill et al., 2019). Their objective is to offer tailored and anticipatory government services and improve their effectiveness. With the wider availablity of large scale data sets, data-driven approaches such as machine learning (ML) applications are on the rise in government policies. In addition to financial and healthcare data, as well as sensor data, governments can leverage large-scale administrative data and merge them with supplementary information on individuals or firms. Governments utilize ML for diverse objectives, such as identifying fraud, predicting crime, managing traffic, and optimizing public services (Ubaldi, Fevre, Petrucci, Marchionni, Biancalana, Hiltunen, Intravaia and Yang, 2019; Organisation for Economic Co-operation and Development, 2019).

Empirical studies on policy effects have focused mainly on causal relationships (Kleinberg et al., 2015). There are vast studies that examine the impacts of subsidized loans or public credit guarantee schemes on sales and employment (Brown and Earle, 2017; Bertoni, Martí and Reverte, 2019; Hottenrott and Richstein, 2020). Causal analysis remains relevant as it allows us to examine the impact of policy interventions. Our study, however, explores

the possibility and direction of policy improvement by taking advantage of the predictive outcomes from the ML model. Using predictions helps to avoid the allocation of public resources to failing firms ("zombie lending," Kwon et al., 2015; Hu and Varas, 2021) and instead directs them towards those who can grow and provide a livelihood to the economy.

ML models provide a practical solution to perform better in various predictive tasks, by estimating functions that perform well in out-of-sample testing (Mullainathan and Spiess, 2017). These models possess highly flexible functional forms to capture complex interactions and nonlinear structures (Varian, 2014). This characteristic of ML has led to a growing body of research beyond engineering that explores its application in resource allocation decision-making for various policy settings.

Andini, Ciani, de Blasio, D'Ignazio and Salvestrini (2018) applied ML on Italy's tax rebate program and showed that the effectiveness of the program could be improved greatly by targeting consumption constrained households. Kleinberg, Lakkaraju, Leskovec, Ludwig and Mullainathan (2018) demonstrated that utilizing ML predictions on crime risk could improve judges' bail decisions. Sansone and Zhu (2021) employed ML algorithms on social security data to identify individuals at risk of long-term income support, offering potential cost savings on government welfare expenses. Andini, Boldrini, Ciani, De Blasio, D'Ignazio and Paladini (2022) showed that ML can increase the effectiveness of public credit guarantee programs by targeting firms that are both creditworthy and credit constrained. They develop two separate ML prediction models for firm credit constraints and firm creditworthiness to assign policy targets that satisfy both conditions.

Our contribution is twofold. First, we demonstrate that leveraging predictive models to target high-expected growth firms can significantly enhance the effectiveness of government credit programs. The predictive outcomes provide valuable information to decision-makers, enabling them to exclude firms with limited growth potential from receiving government credits. When comparing the actual growth rates of sales and assets between firms in the

top 30% predicted sales growth group and the bottom 70%, we found a stark difference in growth rates (27.9% vs. 0.8%). We find that firms ranked in the lowest 30% in terms of predicted growth rate experienced a significant decline in sales after receiving government credits. From a cost-benefit perspective, excluding these firms would result in approximately 30% savings of the programs' budget.

Second, among the various measures of financial constraints suggested in the corporate finance literature, we find that firm size and age are particularly useful indicators of credit constraints. Our paper is related to the literature on firm growth, credit constraints, and government support programs. Previous research has mainly concentrated on evaluating the impact of government programs on firm growth (Fairlie, Karlan and Zinman, 2015; Brown and Earle, 2017; Huergo and Moreno, 2017) or on finding evidence of credit-constrainedness among targeted firms (Zia, 2008; Banerjee and Duflo, 2014; Bach, 2014). Our research differs from previous studies in that we focus on how to factor in firms' credit constraints in actual credit assignments.

Despite all the benefits we present, ML predictions cannot solve all problems and a precaution should be taken (McKenzie and Sansone, 2019). To ensure transparency and reliability for policy implementation, we examine the limitations of ML tools. Understanding these limitations can help the policymakers to address potential issues such as manipulated applications and omitted-payoff, where the former refers to applicant firms fabricating information to secure government credit, while the latter relates to situations where important variables are missing from the model, respectively.

The rest of the paper proceeds as follows. Section 2 describes SME subsidized loan and public guarantee programs in Korea. Section 3 discusses the rationale for targeting credit-constrained firms with high growth potential. Section 4 explains the data used, describes the prediction model, and presents the predicted results. Section 5 presents our findings on applying ML prediction to the subsidized loan and guarantee programs. Section 6 discusses

implementation issues and concludes the article.

2 SME subsidized loans and public guarantees in Korea

Many governments implement policy instruments to help SMEs access financing, enabling their growth and job creation. Credit guarantee schemes have been widely adopted by developed and developing countries since 19th century in Europe to promote private businesses. These schemes are currently available in nearly 100 countries (Green, 2003). Some governments provide direct government loans to SMEs.

In 2018, the Korean government guaranteed 9.7% of all outstanding business loans to SMEs, with an additional 0.6% provided through direct lending. The government's active provision of corporate financing is not limited to Korea alone. For example, the Japanese government provided credit guarantees and direct loans to SMEs, covering 8.3% and 7.9% of the total SME loans, respectively in 2018. Similarly, the Small Business Administration in the United States provided approximately 63,000 loan guarantees, amounting to USD 29 billion, to SMEs in 2018, representing 4.6% of the outstanding business loans to SMEs (Organisation for Economic Co-operation and Development, 2020).

Both public guarantees and subsidized loans aim to aid the growth of companies that have strength in technology and business potential but insufficient financial resources.[1] Public guarantee schemes serve to mitigate the challenges faced by SMEs in accessing credit from the market by transferring the risk from financing institutions to public institutions. Particularly in situations where risks are elevated, the role of public guarantee schemes becomes even more crucial. During the COVID-19 pandemic, for example, numerous countries launched or

[1]The objective is explicitly stated in the annual guidebook for SMEs and Venture Business support programs, which is published by the Korean Ministry of SMEs and Startups (Ministry of SMEs and Startups, 2018).

expanded credit guarantee schemes or direct government loans to support SMEs.[2]

In Korea, subsidized loan rate has varied throughout the study period, and yet it has been consistently 1% to 2.2% lower than market rate between 2012/Q4 and 2017/Q4 in quarter-wise comparison.[3] The interest rate becomes even more favorable for subsidized loan programs targeted at start-up companies, with a reduction of 0.08% to the base rate. The typical loan limit per individual company was set at KRW 4.5 billion in 2015. Public guarantee schemes typically provide guarantees ranging from 85% to 100% of the credit amount for private companies, with a cap set at KRW 3 billion. Subsidized loans are available to SMEs in private sector industries, including manufacturing and services, but they exclude construction, financial services, real estate, and accommodation and food services.

In our case study, the Korean government credit programs include both loan guarantees and subsidized loans, explicitly targeting SMEs that have (1) high growth potential and (2) limited access to finance for their investment. However, assessing the growth potential of applicants poses a challenge, while determining their level of credit constraint is even more intricate. Typically, in the credit application process, the evaluation of each applicant's growth potential relies on a qualitative assessment of the feasibility of their technological and business plans, along with their management skills. The programs do not consider potential availability of loans for a candidate firm from private banks in their decision-making process. However, certain large enterprises, such as listed companies or those with sales exceeding 50B KRW in the previous financial year, are explicitly ineligible.

Firms may apply for government credit support to take advantage of favorable interest rates even if they are not credit-constrained. Another reason may be to sustain their oper-

[2]As of February of 2023, 47 countries have initiated 76 credit guarantee schemes according to data from the World Bank (Map of SME-Support Measures in Response to COVID-19, https://www.worldbank.org/en/data/interactive/2020/04/14/map-of-sme-support-measures-in-response-to-covid-19).

[3]Subsidized loan rates are variable rate decided quarterly, and tied to the policy fund base rate. The policy fund base rate is, in turn, linked to the SME promotion bond procurement rate.

ations when they fail to innovate or adapt to market changes. Coupled with less stringent screening processes, these workarounds can result in substantial inefficiencies within government credit programs. Instead of being allocated to productive firms with genuine credit constraints, government credits may be tied up in zombie firms characterized by stagnant or declining revenues.

The challenge of identifying the policy target can be effectively addressed by utilizing ML applications and employing sales growth as a target outcome variable (dependent variable) for prediction. Instead of relying solely on qualitative evaluations, we demonstrate how the ML approach helps identify credit-constrained firms did we show how to identify firms that are credit constrained? We use the information related to the constrainedness (ccv's) but it doesn't mean that we're looking at 'more constrainted', so to speak. with growth potential by predicting the sales growth rate. Moreover, departing from a passive approach of excluding large corporations as policy targets, we identify suitable indicators to measure the severity of financial constraints, allowing for the allocation of credits to the firms that would receive the greatest benefit.

3 Selecting Suitable Targets for Government Credit Programs

In this section, we develop the theoretical rationale and provide a simulation example of our modeling approach for targeting firms predicted to achieve higher growth compared to other firms while facing credit constraints.

We use the predicted sales growth as a key criterion to choose the beneficiaries of the government credit programs. We posit that the sales growth rate serves as the most appropriate metric since the policy is expected to be more effective when the government provides credit to firms with higher growth potential and greater credit constraints, based on the following

theoretical support.

As in Banerjee and Duflo (2014), we define a firm as credit constrained if its total capital is insufficient to meet its desired amount at the highest interest rate it currently pays. Although we do not directly observe firm-level credit constraints with a specific measure, the level of credit constraints ($w_{i,t}$) is a relative measure of credit insufficiency compared to the firm's demand. As a firm's credit demand is likely to be greater when it expects to grow more, we assume that a firm's credit constraints is positively correlated with the firm's growth ($y_{i,t}$). When we utilize a predictive model to predict a firm's growth, we specifically define this growth as the growth in sales.

Banerjee and Duflo (2014) shows that "if the firm is credit constrained, an expansion of the availability of bank credit will lead to an increase in its total outlay, output and profits, without any change in market borrowing."[4] The logically equivalent contrapositive of this result would be "if an expansion of the availability of bank credit does not lead to an increase in total outlay, output, and profits (without any change in market borrowing), then the firm is not credit constrained." When a firm is not credit-constrained, it would substitute its existing loan with a subsidized loan, and investment will only increase after the refinance.

When government credit is extended to a non-credit-constrained firm, its output remains unchanged, as such firms can already optimize their investments and workforce without additional government credit. In contrast, credit-constrained firms are more likely to experience a higher growth rate when provided with government credit, as it enables them to expand their output with the extra credit available.

We assume that policy impact z is defined to be a linear function of the level of credit constraint and overall potential growth, $z = \alpha_w w + \alpha_y y$, where $\alpha_w, \alpha_y > 0$. The scale factors (α_w, α_y) are determined based on the overall policy objective. The ratio $\frac{\alpha_w}{\alpha_y}$ is larger when

[4] Banerjee and Duflo (2014) presented a theory under the environment where a firm has limited access to cheap bank credit while market borrowings are available at a higher rate. In our case, we can substitute cheap bank credit with government-subsidized credit.

the policy places greater importance on providing credits to more credit-constrained firms compared to higher-growth firms. The policy decision is to determine its eligibility based on the firm's potential growth, $\Omega_c = \{(w, y) : y \in \mathcal{S}_c\}$, where $\mathcal{S}_c$ is the eligibility condition set by c.

Definition 1. *(Policy effectiveness) The policy effectiveness of a policy c is defined as*

$$\mu(c) = E(Z|Y \in \mathcal{S}_c) = \int_{\Omega_c} z dP_{w,y|y\in\mathcal{S}_c},$$

where $P_{w,|y\in\mathcal{S}_c}$ is conditional distribution of (W, Y) given $Y \in \mathcal{S}_c$.

Definition 2. *(Policy superiority) For two policies $c \neq c'$ with $P_{w,y}(\Omega_c) = P_{w,y}(\Omega_{c'}) = \pi > 0$, c is a more effective policy than c' when $\int_{\Omega_c} z dP_{w,y|y\in\mathcal{S}_c} > \int_{\Omega_{c'}} z dP_{w,y|y\in\mathcal{S}_{c'}}$.*

Note that $\pi > 0$ implies that, at the decision-making stage, it has already been determined that a certain portion of the firms will be granted credits. Now we examine the case where $\mathcal{S}_c = \{y : y > c\}$. That is, the policy eligible set is determined by considering firms with growth potential that exceeds the policy threshold, c. We assume a simplified situation where two random variables, W and Y follow a bivariate normal $(W, Y) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We can denote the mean vector and covariance matrix as $\boldsymbol{\mu} = (\mu_W, \mu_Y)$, and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_W^2 & \rho \\ \rho & \sigma_Y^2 \end{pmatrix}$. We assumed a positive correlation between a firm's credit constraints and its growth, meaning $\rho > 0$.

Proposition 1. *Consider a policy that restricts its eligibility to firms whose potential growth is larger than the threshold, c. Under the bivariate normal distribution assumption with a positive correlation, the expected value of policy impact is greater when its eligibility is limited to Ω_c compared to the policy without any eligibility restrictions.*

Proof. Observe that $E[Z|Y > c] = \alpha_w E[W|Y > c] + \alpha_y E[Y|Y > c]$. Given the properties of the normal distribution, the conditional distribution of W given Y follows a normal distribution, $W|Y \sim N(\mu_*, \sigma_*^2)$, where $\mu_* = \mu_W + \rho\frac{\sigma_W}{\sigma_Y}(Y - \mu_Y)$ and $\sigma_*^2 = (1 - \rho^2)\sigma_W^2$.

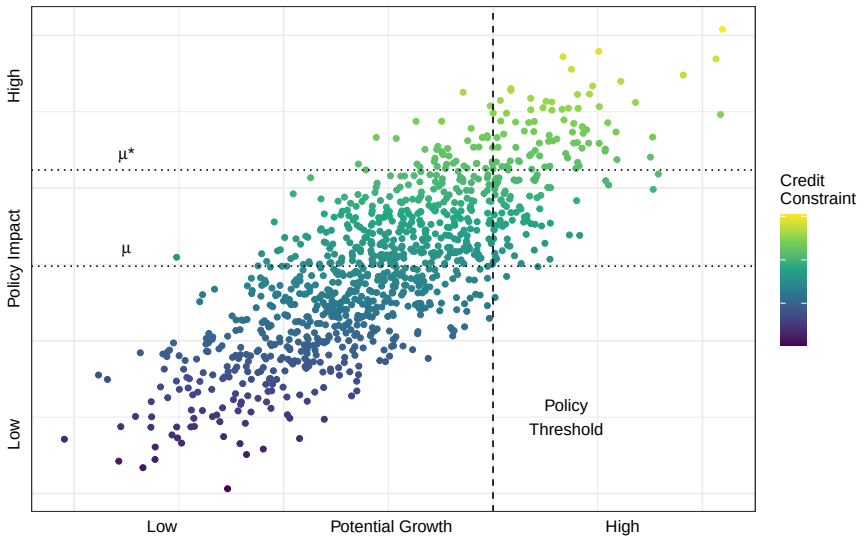

Figure 1: A Simulated Example: Policy Impact on Firms Distributed by Predicted Growth and Credit Constraints

When we introduce the constraint $Y > c$, we apply the properties of a truncated normal distribution to find that $\mathrm{E}(W|Y > c) = \mu_W + \rho\frac{\sigma_W}{\sigma_Y}(E[Y|Y > c] - \mu_Y)$. Here, $E[Y|Y > c] = \mu_Y + \sigma_Y\phi(c;\mu_Y,\sigma_Y)/[1 - \Phi(c;\mu_Y,\sigma_Y)]$, where $\phi(\cdot;\mu_Y,\sigma_Y)$ and $\Phi(\cdot;\mu_Y,\sigma_Y)$ represent the probability density function and cumulative distribution function of the normal distribution with mean μ_Y and standard deviation σ_Y, respectively. Since $E[Y|Y > c] \geq E[Y]$, $E[Z|Y > c] \geq E[Z]$. □

Figure 1 shows a hypothetical distribution of applicant firms by their predicted growth on the x-axis and the level of credit constraints on the y-axis. In particular, we assumed that the potential growth rates and the level of credit constraints are jointly normally distributed with a positive correlation between the two variables. [5] Credit constraints can be a source of distortions that change the marginal product of capital or labor. Every dot represents a potential applicant firm with $(cc_{i,t}, y_{i,t})$ for government credit. Policy impacts are captured

[5]The assumption on the distribution can be strong. Recent empirical studies show that firm growth rates follow a Laplace distribution, a tent-shaped form that has a peak around the center and fatter tails.(Bottazzi and Secchi, 2006; Arata, 2019) The mechanism illustrated here goes through when we assume a Laplace distribution for firm growth rates.

by the color density of dots where darker color denotes lower impact. The solid line is the policy threshold where credit is provided to the firms whose expected growth falls on the east side of the threshold line.

The proposition demonstrated that, contingent upon specific conditions, the expected policy impact increases with the adoption of an eligibility criterion that extends credit to firms whose projected growth exceeds a predefined threshold. In Figure 1, the expected policy impact with a threshold, $\mu_*(c^*)$ is higher than the average policy impact without one, μ.

However, raising the threshold c^* to maximize the expected policy impact $\mu_(c^*)$ comes at the cost of reducing the number of firms that receive government credit. Typically, government credit programs operate within budgetary constraints for each year. Given these limitations, there are several approaches to select eligible firms. One option is to rank firms based on their projected growth rates, allowing the program to fund the highest-ranked applicants within the budget. Alternatively, the program can establish a specific growth rate threshold, such as zero or the average industry growth rate, to limit eligibility. Recognizing that predictions are not always perfectly accurate, the program could also implement an additional review process for firms whose projected growth rates fall below the threshold, rather than outright disqualifying them from consideration.

Another reason for targeting the sales growth rate is that it serves as a key performance indicator (KPI) and can be utilized to evaluate the impact of government support. The Korean Ministry of SMEs and Startups has explicitly designated sales growth rates as the primary performance indicators in their annual performance and accountability report.[6] The National Finance Act of Korea mandates that each ministry submits a performance plan for the upcoming year's budget and a performance report for the previous year's budget to the Minister of Strategy and Finance (National Assembly of the Republic of Korea, 2023). These

[6]Cite: Small and Administration (2016) "Fiscal Year 2015 Settlement of Revenue and Expenditures" by the Small and Medium Business Administration (currently the Ministry of SMEs and Startups).

reports include program goals and performance indicators that are established in advance, and the actual performance is reported for the purpose of financial management.[7] By targeting the sales growth rate, a what-if analysis can be performed to assess the effectiveness of the policy. This approach can assist policymakers in identifying suitable targets from the pool of candidates.

4 Data and modeling methodology

4.1 Data

Our model builds on the two data sources. The first source is the SIMS database, an extensive Korean administrative database that provides yearly information on subsidies given to SMEs at different government levels. However, this database lacks important details such as balance sheet information and basic characteristics of the SMEs. To address this limitation, we complement the SIMS data with information from the Korea Enterprise Data (KED) database. The KED is the largest database in South Korea that specializes in providing credit information on SMEs.[8]

The SIMS database keeps track of information on project names, functional area, department in charge, budgets, and their support history at the firm level from 2010.[9] The KED database provides information on revenue, assets, profits, capital, liability, industry, establishment year, and corporate forms of business ownership. We link the yearly firm-level data from SIMS with KED database using the business registration number.

Our study focuses on the firms that received government-guaranteed loans and direct

[7]The growth in total assets can also be considered a predicted variable given that loan is mainly made for financing major fixed assets. This study presents the growth in total assets based on the predicted sales growth rate in Section 5.

[8]see http://www.kodata.co.kr/en/ENINT01R1.do

[9]Korean government introduced "Integrated Management System for Small and Medium Enterprise Aid Programs (SIMS)" to improve the efficiency of SME support policies. It is a database platform that integrates and manages information on SME support projects implemented by any central and local governments.

government loans between 2010 and 2015. There are four national programs that share similar objectives in providing financing to SMEs with growth potential. Two of these programs are credit guarantee schemes: the Credit Guarantee Support program by the Korea Credit Guarantee Fund and the Technology Credit Guarantee program by the Korea Technology Finance Corporation. The remaining two are government loans. One is from the Start-up Company Support Fund, a direct loan program to support start-up companies by the Korea SMEs and Startups Agency. The other is On-lending, an indirect loan program by the Korea Development Bank. These four programs provided guarantees or loans to over 80,000 firms annually during the period from 2010 to 2015. Using the linked data between SIMS and KED, we identify firms that received loans from any of the four major programs and treat them as a subsidized group.

We further subdivide the sample of firms based on their age, specifically focusing on start-up companies that are up to 6 years old. Age is defined as the number of years since the firm's establishment. In Korea, SME support projects are classified into start-ups (less than 7 years old) and non-start-ups (7 years and older). Various programs, including the Start-up Company Support Fund, exclusively target start-up companies. By focusing on start-up companies, we can consider the different growth patterns that exist across different age groups. Research has shown that young firms exhibit much higher mean and dispersion of growth rates compared to older firms, conditional on their survival (Decker et al., 2014; Kim, 2017).

Next, to address potential data contamination issues, we excluded firms that received subsidies from government credit support policies other than the four programs during the period from 2010 to 2015. Our data allow us to control for any influence from other government programs. After this exclusion, we are left with two distinct groups in a given year: the subsidized group, consisting of firms that received their first government credit support during the same period, and the non-subsidized group, comprising firms that did not receive

any government credit support.

Based on this dataset, we build separate datasets for each three-digit industry and year pair since we estimate ML model for each pair. Even though there is no fixed requirement for the minimum number of data for ML, we exclude any pair which has fewer than 500 firms to ensure that a model has enough data to learn patterns from covariates. We also exclude companies whose balance-sheet information is missing from the previous year. We apply the eligibility requirement for government loans to SMEs and exclude companies with sales of KRW 50 billion or more or total assets of more than KRW 100 billion in the previous year. Companies listed on the Korea Exchange (KRX) or the KOSDAQ market, cooperatives, and non-profit organizations are excluded by the same requirement.

After applying the aforementioned criteria, we narrowed down the dataset to 395,541 firm observations for analysis from 2010 to 2015. During this period, approximately 17.6% of the firms received their first government-subsidized loan. We believe that the data is suitable for analysis as it covers the majority of the SMEs in Korea and includes all government credit support provided to SMEs.

Table 1 shows the summary statistics for variables used in the analysis. The analysis specifically presents the characteristics of subsidized firms during the initial year of the subsidy and those of non-subsidized firms for the corresponding year. This allows for a comparison of the specific characteristics of subsidized firms at the start of their subsidy period and non-subsidized firms at a corresponding point in time.

On average, sales and profits are comparable between the subsidized firms and not-subsidized firms, whereas total assets, tangible assets, and liabilities are significantly lower for subsidized firms. The subsidized firms tend to be younger and have more employees, suggesting that they operate with relatively limited capital, potentially due to financial constraints. These firms also exhibit a higher debt-to-capital ratio and lower cash flow, indicating that they may face greater credit constraints. The combination of their younger age

and smaller asset size further suggests a higher likelihood of experiencing such constraints.

Subsidized firms tend to grow at a faster rate in both sales and total assets than their counterpart. It is possible that this difference in growth rates can be attributed to the effects of the subsidy, or to their demographic characteristics, such as being young and small.

Table 1: Summary Statistics

	Subsidized		Not Subsidized	
Variable	Mean	SD	Mean	SD
Sales (1M)	3399	(8189)	3295	(7185)
Profit (1M)	141	(635)	139	(569)
Total assets (1M)	1620	(6335)	2070	(5091)
Tangible assets (1M)	549	(2851)	613	(2280)
Liability (1M)	1115	(3721)	1259	(3042)
Paid-in Capital (1M)	223	(701)	326	(811)
Subsidized Loan (1M)	368	(733)	-	-
Number of employees	9.88	(21.17)	8.23	(17.52)
Age	2.79	(1.64)	3.58	(1.60)
Debt-to-capital ratio	0.68	(0.40)	0.65	(0.56)
Cash flow ratio	38.00	(2724.79)	96.72	(1636.74)
Markup	1.04	(0.70)	1.06	(1.35)
Sales growth rate	0.10	(0.60)	0.07	(0.55)
Asset growth rate	0.21	(0.46)	0.19	(0.43)
Number of firms	69,678	-	325,863	-

Note: The table reports the mean (and std. dev. in parentheses) of each variable for subsidized and not subsidized firms. The number of observations is reported at the bottom row. Variables with (1M) are reported in current million KRW. The debt-to-capital ratio is defined at the firm level as the ratio of total liabilities to total capital. Total capital is calculated as the sum of liabilities and total equity. Cash flow ratio is operating income over lagged tangible assets.

4.2 Predictive model

ML models can be useful when the (potentially nonlinear) relationships between the target variable and many predictor variables are to be learned (Varian, 2014). As implied by its name, an ML model has the capability to automatically "learn" such relationships, which becomes particularly valuable in cases where the volume of data is too vast for human decision makers to conduct comprehensive analysis and investigation.

Our modeling approach aims to *predict* the performance of a firm in the future, given the information available at the current year. If there exist critical factors that affects the firm's

future growth, the model should be able to reflect the underlying structure by including the effect of such factors into the modeling structure. Our approach directly addresses the problem from decision making perspective: making a future prediction with the information available at the moment when the subsidy beneficiary is selected.

Consider the following predictive model

$$y_{i,t} = f(\mathbf{x}_{i,t-1}) + \epsilon_{it}, \tag{1}$$

where y_{it} is the sales growth for firm i at year t, $\mathbf{x}_{i,t-1}$ the firm characteristic available at $t-1$, and ϵ_{it} is the independent and identical error term with mean 0 and unknown variance σ^2. Firm i's sales growth $(y_{i,t})$ is defined as the log difference of sales of a firm $(S_{i,t})$, that is, $y_{i,t} := \log S_{i,t+1} - \log S_{i,t}$. One can view (1) as estimating a conditional expectation of the growth rate in the following year, given $\mathbf{x}_{i,t-1}$, the set of characteristics of the current year. The functional form $f(\cdot)$ includes a wide class of supervised learning models, such as (generalized) linear models, additive models, or regression tree models (Hastie et al., 2009).

Though the model in (1) can be estimated in an automated manner, its execution still needs a thorough pre-processing. First, we create blocks of firms divided by industry sectors and years, and a separate model is trained for different blocks. Different industries may have very different growth structure (e.g., brick-and-mortar versus technology firms), and exhibit distinct annual growth rates due to variations in sector-specific factors such as consumer demand and the business cycle. Such disparities can be incorporated by including the blocks as input variables, but building a separate model allows more flexible modeling structure to incorporate dissimilarity between industries. The industry is defined by the three-digit industry classification and we only considered manufacturing and service industries groups with more than 500 firms. The number of industries varies from 42 to 46 each year. The actual sales growth rate was ranked among firms within each group and year, to further

stabilize the variation between firms.

Second, we carefully separate the model training process from the evaluation samples by strictly choosing the firms in the next year as a hold-out sample. For example, after estimating the model for a group of firms belonging to the C20 industry (Manufacture of chemicals and chemical products except pharmaceuticals) in using the data up to 2014, the model prediction was made for the firms in the same industry in 2015, and its accuracy was evaluated. The reasoning behind this separation process is twofold. First, it naturally tests the practical applicability of the machine learning method; the decision for the funding for the fiscal year 2024 needs a prediction that was made based on the information available in the year 2023. Second, it prevents the information in the same year from leaking into the predictive model. If the model is built using the data strictly up to 2022 to make the prediction for 2023's performance, potential data contamination can be minimized.

Third, the initial data set is built as an annual data set consisting of sales growth, revenue, assets, profits, equity, net income, venture company status, industry, number of full-time employees, age, amount of subsidized loan, and corporate forms of business ownership [10] for every year. The data set is re-structured so that for every annual sales growth, the input variables contain information on firms up to the past two years of their application. In the initial data wrangling process, input variables that have more than 20% missing were excluded from the analysis.

Lastly, we screen the firms within each industry block to choose only a subset of the firms that are similar to the firms that were subsidized. In the process, five nearest neighbors among those who were not subsidized in the input space are chosen for each subsidized firms. After the five nearest neighbors are selected for all the firms in this block, the union of the selected firms, after being joined with the subsidized firms, is used as the data set for this

[10] listed on KOSPI (Korea Composite Stock Price Index) or KOSDAQ (Korean Securities Dealers Automated Quotations)/corporations/Sole proprietorship/limited liability company/general partnerships/limited partnership/foreign corporation

block. The majority of the firms did not get any subsidy, so this process reduces the number of total firms in the modeling and validation process. This matching process makes the modeling more difficult as it both reduces the sample size and makes the firms more difficult to distinguish, and yet it ensures that the prediction model is estimated for firms that share similar characteristics with the applicants. In presenting our results, we also show outcomes of non-subsidized firms for comparison purposes. Utilizing matching techniques facilitates more fair comparison because it excludes the firms that have too different characteristics to be compared with the subsidized firms.[11]

Amongst many possible modeling choices, we use random forest (RF) as our prediction model. As its name "forest" implies, the methodology is built based on many small regression tree models. A regression tree model partitions an input space into a set of intervals and produces a Cartesian product of those partitions. Prediction from a regression tree model for a given $\mathbf{x}$ can be expressed as

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} \hat{c}_m \mathbb{I}(\mathbf{x} \in R_m), \tag{2}$$

where $R_1, \cdots, R_M$ are the partition of the input space of $\mathbf{x} = (X_1, \cdots, X_p)$ into M regions, $R_1, \cdots, R_M$, and $\hat{c}_m$ is the sample mean of y_i for $i \in R_m$. The algorithm finds the partition until its accuracy does not improve.

RF is an ensemble method to further improve prediction accuracy by combining predictions from multiple decision tree models. The main idea is based on the law of large numbers – average of multiple samples is less volatile than a simple sample. In each iterate, RF randomly selects a subset of predictors, a bootstrap sample, and builds a tree model using this subset. Since this small tree model only uses a subset of the predictors, it may

[11] By using the matching procedure, we address the issue of selection bias and endogeneity that may arise due to the relationship between the characteristics of credit applicants and the actual outcome. However, we do not view this problem as a significant concern since our prediction model will only be applied to the applicants.

not be very accurate ('weak'). RF builds many weak trees, and makes a prediction for the new target data from each tree, and averages them to make final prediction. Variance of the final prediction tend to be reduced by aggregating predictions from multiple models. This reduction is due to weak correlations between different trees, because different trees use different different predictors. During the training process, the algorithm further partitions the bootstrap sample into training and testing data, where training data is used to fit the model and testing data is to tune the model. 63.2 percent of the firms are part of the training sample while the remainder belongs to the testing sample. The number of trees was set to be 500.

All the empirical analysis in this work was conducted with R (R Core Team, 2022), where data wrangling process and computation used `tidyverse` (Wickham et al., 2019) and `randomForest` (Liaw and Wiener, 2002) packages, respectively.

4.3 Prediction results

Model's performance is assessed by its prediction accuracy for firms' sales growth. To this end, Figure 2 presents the actual average growth rate of sales for both subsidized and not subsidized firms across 10% percentiles of predicted growth. The percentiles are calculated for each year and industry sector. It is clear that the actual sales growth rates tend to be higher for the firms whose predicted growth is high. The highest 10% of subsidized firms in terms of predicted growth rate have an average annual sales growth rate of approximately 50%, while for the next 10% around 19%. On the contrary, those predicted to belong to the three lowest groups show negative growth in each group, on average. Subsidized firms predicted to be in the top performing group showed slightly higher growth rates compared to non-subsidized firms, but its difference is not substantial.

The findings demonstrate the efficacy of ML can in identifying firms that are likely to perform well in the following year. The predictive outcomes can provide insights for

decision-makers to rule out firms with limited growth potential from consideration. It is still important to exercise caution when interpreting these results, as they do not imply a causal effect of the subsidized loan. Various other factors contribute differences in growth among firms. For instance, younger firms tend to grow faster, and subsidized firms are generally younger. However, we meticulously constructed data sets so that the control group (non-subsidized firms) closely resembled the treatment group (subsidized firms) in terms of firm demographics. This approach can filter out the impact of observable factors. Thus, the results suggest that either a significant portion of subsidies was allocated to firms with limited growth prospects, or that the subsidy itself made only a modest contribution to the growth of the firms.

While it is not uncommon for certain firms to experience negative sales growth in a given year, providing subsidies to firms projected to decline would be a hard sell to taxpayers. The predicted KPIs can serve as a valuable tool for policy decision-makers to prevent such outcomes. Government officials can use the predictions as reference indicators. For example, they can reconsider or reexamine providing loans to firms in the bottom 30% of the performance prediction group. Alternatively, loans can be reallocated based on predicted performance groups, ensuring that firms in the lower groups receive less funding overall. In the subsequent Section 5, we explore the quantitative benefits of utilizing ML predictions in a government loan program.

5 Effectiveness of ML predictions and human decision

In this section, we delve into the utilization of the predicted information during the actual policy stage to select the loan recipients. We propose the implementation of a straightforward decision rule: granting a loan to a firm only if its predicted sales growth rate ranks among

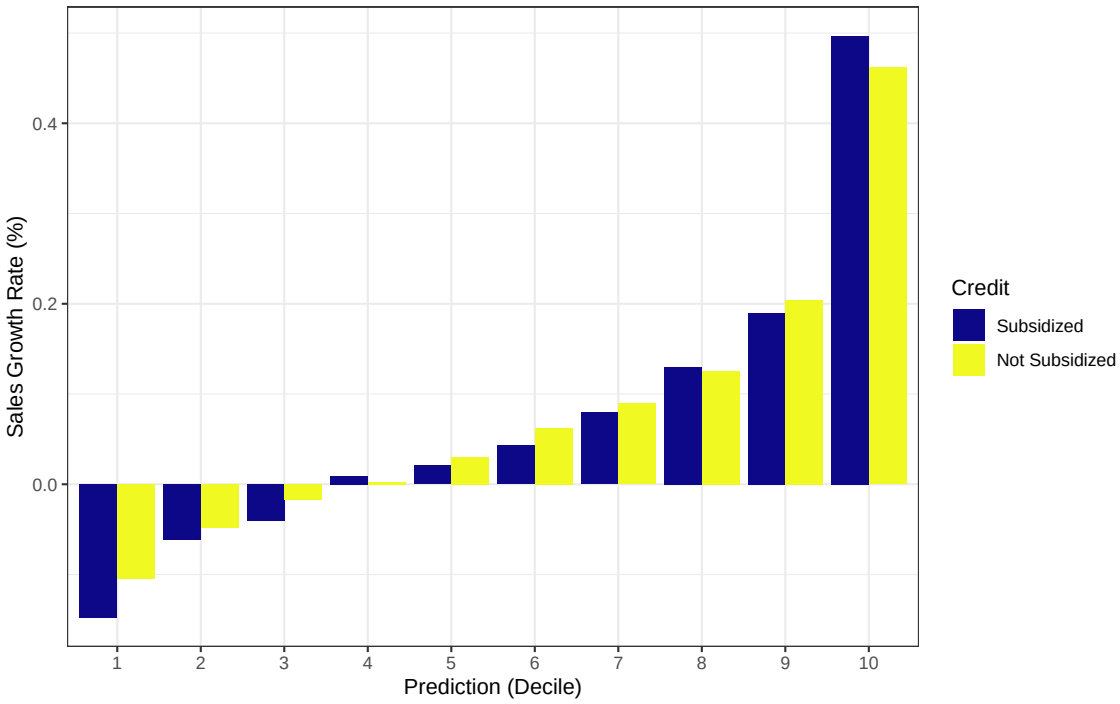

Figure 2: Machine Learning Predictions and Actual Outcomes
Notes: The figure shows the average growth rate of sales for both subsidized firms and not subsidized firms across deciles of predicted growth.

the top 30% of all applicants. It is important to note that the 30% threshold is used here as an illustrative example, and the actual threshold can be determined based on factors such as the available size of the subsidy, or the desired level of targeted growth for the policy. Besides the fact that most support projects set the recipients' sales growth rate as their performance indicator, we prefer this measure since it is reasonable to expect firms to grow their sales when they are credit constrained with their growth prospect. Moreover, sales is a direct performance indicator that has less room for a firm to manipulate compared to other measures such as operating profit. We apply the simple allocation rule providing loans to the top 30% of predicted sales growth in the following section 5.1.

Another important dimension to consider when selecting loan recipients is that their level of credit constrainedness. SMEs can be credit constrained either because they have a short credit history (young) or because they lack enough collateral (small). When firms are not credit constrained, they will increase their production or investment after they substitute subsidized credit for market credit since subsidized credit is cheaper than market credit (Banerjee and Duflo, 2014). The sales will increase more when the subsidy is made to more credit-constrained firms. Thus, the effectiveness of a program depends not only on selecting firms with the highest expected growth but also on providing credit to more credit-constrained firms. In section 5.2, we consider the proxy variables for credit constrainedness combined with predicted sales growth rate.

5.1 Utilizing predictions from machine learning to find targets

We quantitatively assess the impact of machine learning on enhancing the efficiency of subsidized loans. We identify potential recipients using ML predictions model in Section 4.2. We compare the average actual growth rate of sales and assets of firms who belong to the top 30% predicted sales growth vs the bottom 70%. The sales growth rate is a metric that represents a company's performance, while the growth rate in total assets indicates changes

in a company's investments. We calculate the growth rate by taking the difference between the logarithm of sales or total assets in the current year and the next year. We then average these growth rates over each year from 2011-2015. A.I in the appendix A shows the average growth rates for each year during the period.

The average actual sales growth rates and total asset growth rates of firms in the top 30% were significantly higher compared to those in the bottom 70%, as determined by the predicted sales growth rate. Figure 3 shows the results for firms that received subsidized loans and other firms that did not get any loans. Among the recipients of subsidized loans, the sales grew 27.9% on average for the top 30% firms while it grew negatively by 0.8% for the bottom 70% firms (Fig. 3 panel (a)). The difference in the growth rate is stark between the two groups. On average, 70% of subsidized loans could not help firms grow in sales. In addition, for firms in the lowest 30% of predicted growth rate, there was a significant decrease in sales after loan disbursement. In terms of investment, the top 30% of firms showed a higher increase compared to the bottom 70% group (24.5% vs 19.7%) although the difference was not as stark as observed in the case of sales growth rate(Fig. 3 panel (b)). While the bottom 70% group exhibited an average increase of 19.7% in their investment, however, their sales growth remained close to zero.

Figure 3 also shows firm performance among not subsidized firms. The average sales increased by 25.3% for the top 30% group while it decreased by 0.1% for the bottom 70% group. The comparison of growth rates between the subsidized firms and not subsidized firms may not be reasonable since the subsidized firms can be a selected group of firms that has different characteristics than other firms. They can be more credit constrained or have a higher incentive to substitute subsidized credit for their existing credit. Thus, we do not infer any causal effects of getting subsidized loans on the firm performance. However, it is noteworthy that the sales growth rate of subsidized firms was higher than that of non-subsidized firms in the top 30% category, while it was lower among the subsidized firms in

the bottom 70% category. Additionally, subsidized firms exhibited greater growth in assets than non-subsidized firms across both predicted sales groups.

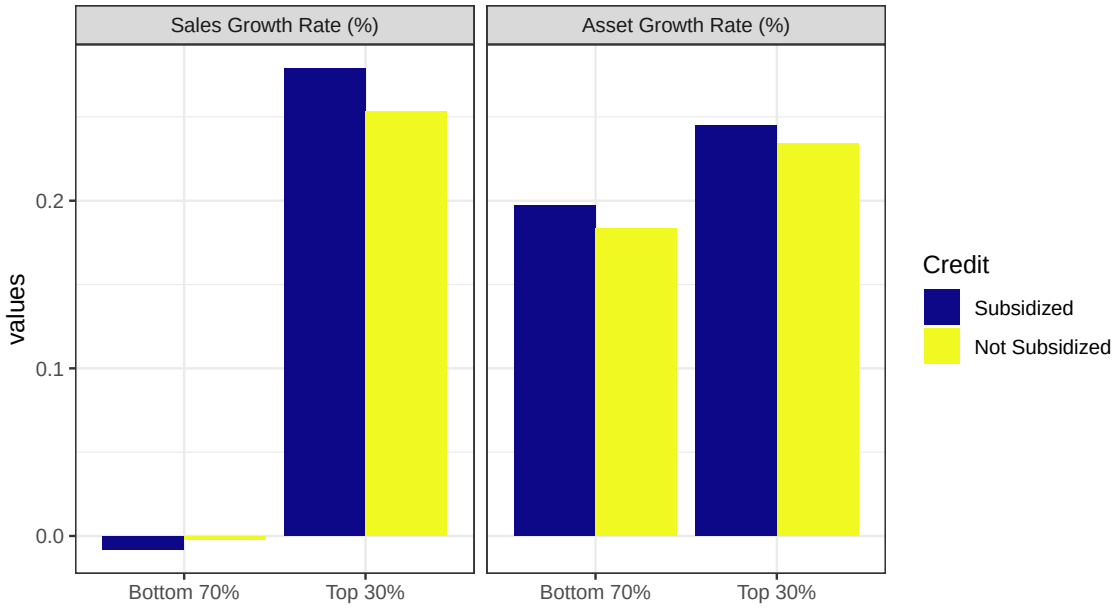

Figure 3: Firm performance comparison by ML prediction target
Notes: Panels (a) and (b) display the average growth rates of sales and total assets, respectively, for firms in the top 30% and bottom 70% based on the predicted sales growth. The average values are presented for both subsidized and non-subsidized firms in all panels.

5.2 Considering credit constrained firms

Once we substitute the subjective evaluation on the applicant firms with their predicted sales growth, we can explore ways to consider credit-constrainedness of the firms when selecting loan recipients. Our study employs four measures suggested in the corporate finance literature to access the financial constraints of the firms, which are firm size, age, debt-to-capital ratio and cash flow ratio. This analysis can shed light on the potential improvement of resource allocation. By combining the outcomes of the predictive model with proxy information on credit constraints, we can identify firms that would benefit the most from receiving credit support.

Various methods were proposed to measure the extent of a firm's credit-contrainedness. Kaplan and Zingales (1997) utilized qualitative data from financial managers of firms to measure the level of financial constraints. By applying the regression coefficients of Kaplan and Zingales (1997), Lamont et al. (2001) estimated the degree of financial constraints as an index constructed from a linear combination of five accounting ratios, such as cash flow to total capital and debt to total capital. In addition, Whited and Wu (2006) created an index of a firm's finance constraints using the generalized method of moments (GMM) estimation of a structural model. These studies relied on variables derived from accounting information to estimate a firm's degree of financial constraints. These variables included operating profit or cash flow over tangible assets, Tobin's Q, debt-to-capital ratio, dividends, asset size, firm or industry sales growth rate.

However, when Hadlock and Pierce (2010) revisited the issue of measuring credit-constrainedness using new data, they discovered several variables that lacked consistency. They highlighted that, after adjusting for size and age, only two variables (a firm's leverage and cash flow) reliably predicted the firm's constraint status. They further argued that these two variables had limitations due to endogeneity problems, especially for the leverage ratio, which could result in biased estimations. As a result, they suggested that a firm's asset size and age were useful proxies for the degree of financial constraints since they demonstrated consistency and were less susceptible to endogeneity problems.

In accordance with Hadlock and Pierce (2010)'s suggestion, we utilize a firm's asset size and age as the primary proxies for credit-constrainedness because they are comparatively exogenous. We anticipate that a company will be more credit-constrained if it is smaller (with fewer assets available for collateral) and younger (with less credit history). In addition, we investigate a firm's debt-to-capital ratio and cash flow ratio as they have consistently been shown to predict a firm's constraint status in the study by Hadlock and Pierce (2010). For each year and industry group, firms are divided into four quartiles based on each of the four

measures of constraints, and the average actual growth rate of sales and assets is compared to determine whether the growth rates are higher for more credit-constrained firms.

Figure 4 displays the average growth rate of sales and assets for subsidized and not subsidized firms by the quartiles of total assets and ML predicted growth. The results indicate that, for subsidized and not-subsidized firms, the smaller the total assets, the higher the sales growth rate. The findings cannot be solely ascribed to the base effect, as the observed pattern does not persist for firms that have been in business for more than seven years.

Furthermore, the observed negative association between sales growth rate and firm size was particularly pronounced for the top 30% of firms ranked by predicted sales growth and belonging to the subsidized group. Firm size has a significant impact on firms with high growth potential. In the high-growth expected group, subsidized firms exhibited higher sales growth rates than non-subsidized firms across all firm size quintiles. The expectation was that providing support to firms facing credit constraints would lead to an increase in their sales growth rates.

Regarding the growth rate of assets, the subsidized firms experienced higher growth rates compared to not-subsidized firms. However, the gap between the high predicted growth (top 30%) and low predicted growth (bottom 70%) groups in terms of asset growth rate was much smaller than that observed for sales growth rate. Notably, while the smallest size group in the low predicted growth category demonstrated a higher growth rate in assets (33.4%) to the same-sized group in the high predicted growth category (31.1%), their sales growth rate was markedly lower at 1.8% compared to the 28.8% of high predicted growth firms.

In addition to the size of a firm's assets, age serves as a valuable indicator of a firm's level of credit constraint. A firm's age is associated with its credit history, credit rating, and management expertise, all of which can hinder younger firms from obtaining the necessary funding.

Figure 5 presents a comparison of the average sales and asset growth rates across age and

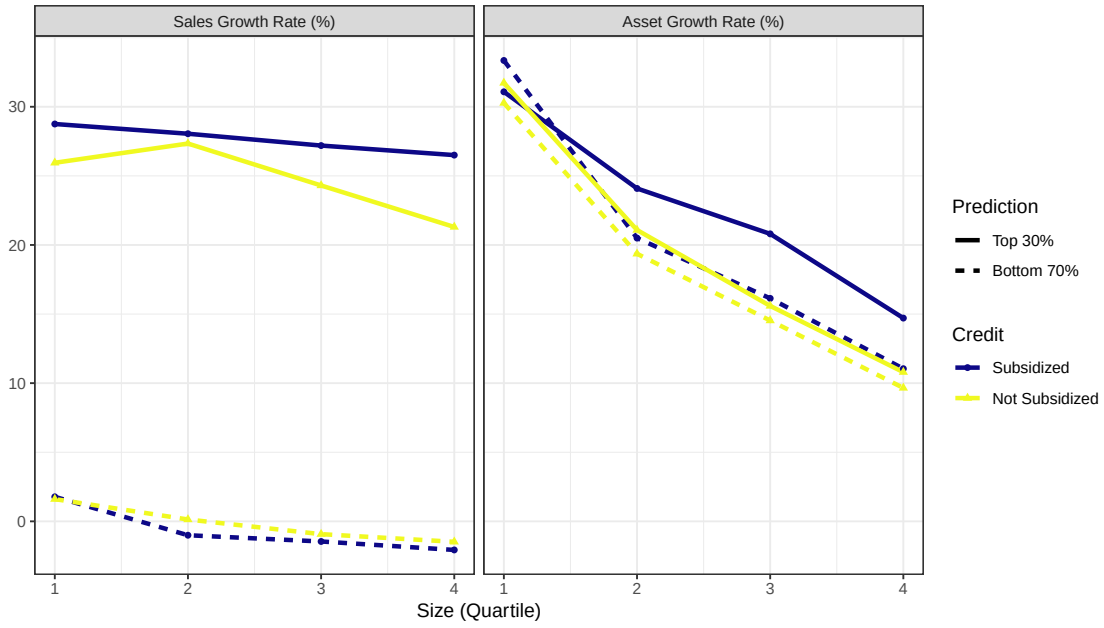

Figure 4: Firm performance comparison by ML prediction and size
Notes: The figure displays the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target and asset size quantiles defined at each year and industry group.

ML predicted growth categories. Younger firms exhibit higher growth rates of both sales and assets. It is important to note that the sample includes firms that are less than 7 years old, meaning that the oldest firms have a business history of 6 years. These firms are relatively young and have qualified for subsidy loans for start-ups. In the high-growth expected group, the sales growth rate of subsidized firms was higher than that of not-subsidized firms for the youngest firms but lower across all other age groups. The sales growth rate of subsidized firms was higher than not-subsidized firms for all age groups in the low-growth expected group.

Remarkably, the growth rate decreases sharply as we move across age, particularly for the subsidized firms with high expected growth. For instance, the oldest group showed 12.2% growth, despite being predicted to be in the top 30%, whereas the same age group of non-subsidized firms exhibited an average growth rate of 16.9%. These results indirectly suggest

that relatively older firms are less constrained when it comes to securing funding for their desired investments. This observation becomes more apparent when we consider both the size and age of firms together. Typically, younger firms are small and grow their asset size as they mature, thereby creating a negative correlation between age and firm size (see Clementi and Hopenhayn (2006)). However, this relationship exhibits variation since not all firms start out small and some firms remain small.

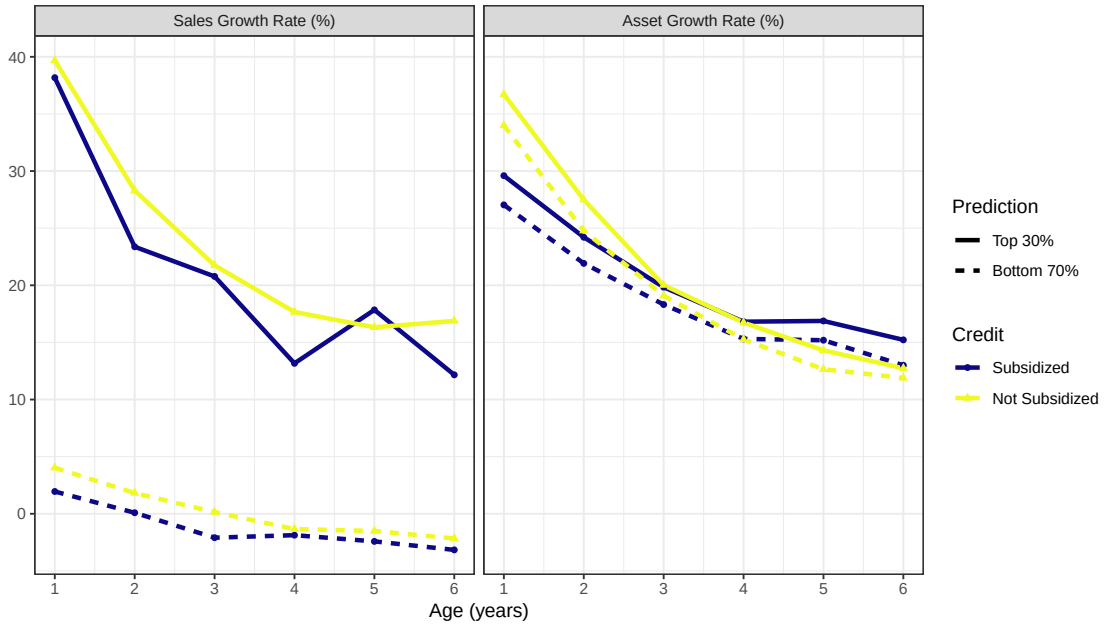

Figure 5: Firm performance comparison by ML prediction and age.
Notes: The figure presents the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target and age (defined at each year).

We further divided firms into size and age and compared the growth rates in Figure 6. Our analysis revealed that the decline in sales growth rate among subsidized firms with high expected growth is particularly evident among the smallest size group. For the smallest size group, we observed that the negative association between growth rate and age is significantly more pronounced among subsidized firms than among non-subsidized firms. The analysis in Figure 4 showed that, on average, smaller firms exhibit higher growth rates. However, this may not be the case for relatively old firms that have remained small. We can infer that they

are not as credit-constrained, even though they received subsidized loans.

There is ample room for efficiency improvement in the subsidy loan program if firm age is taken into account further than just using it as an eligibility criterion. Program managers would be better to pay extra caution on relatively old but small firms. Our analysis emphasizes the importance of understanding the role of age and size in growth dynamics for effective policy-making. Haltiwanger et al. (2013) found that small, mature businesses contributed negatively to job creation, while young firms are significant job creators. Small and mature businesses may have limited motivation to grow or innovate as found in Hurst and Pugsley (2012).

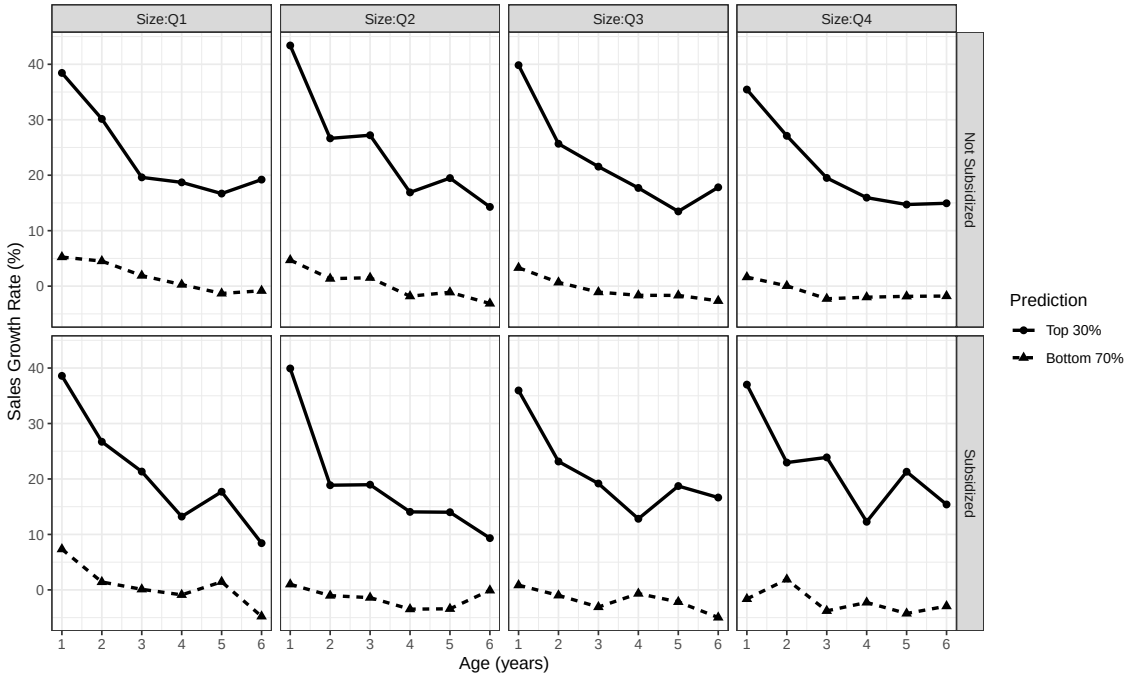

Figure 6: Firm performance comparison by ML prediction, size, and age.
Notes: The figure shows the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target, age, and asset size quantiles defined at each year and industry group.

Now, we explore the debt-to-capital ratio and cash flow ratio as measures of the credit-constrainedness to identify potential policy targets. In models predicting credit constraints,

as observed in studies such as Lamont et al. (2001), Whited and Wu (2006), and Hadlock and Pierce (2010), the debt-to-capital ratio displayed a positive association with an index of credit constraints. Conversely, the cash flow ratio exhibited a negative relationship with the index of credit constraints. In Section 4, we observed that subsidized firms exhibited a higher debt-to-capital ratio and lower cash flow compared to not subsidized firms. Both indicators point in the direction that subsidized firms encounter greater credit constraints.

Figure 7 and 8 present a comparison of the average sales and asset growth rates based on the ML-predicted growth categories for each quartile of the debt-to-capital ratio and cash flow ratio, respectively. The results reveal several noteworthy observations. First, both measures exhibit an association with credit constraints, aligning with the findings of previous studies mentioned earlier. Specifically, higher quartiles of the debt-to-capital ratio and lower quartiles of the cash flow ratio are associated with higher sales growth rates. Additionally, a notable jump in the sales growth rate is observed among firms in the highest quartile of the debt-to-capital ratio and the lowest quartile of the cash flow ratio. It is important to note that this relationship is primarily evident among firms ranking in the top 30% of the predicted sales growth rate.

Second, more credit-constrained firms in these two measures, as indicated by higher quartiles of the debt-to-capital ratio and lower quartiles of the cash flow ratio, may experience faster growth in sales but show a smaller increase in assets. Thus, unlike the case of age and size, the sales growth rates and asset growth rates do not move in the same direction as we move along the level of credit constrainedness in these measures. There are a couple of reasons that can explain this phenomenon.

Firstly, companies relying more on debt financing or having a low cash flow ratio may face higher operating expenses or significant financial obligations, such as debt payments or interest expenses. These obligations can limit the amount of cash flow generated from their tangible assets, affecting their ability to invest in growth opportunities or increase their asset

base.

Secondly, credit-constrained firms may prioritize the allocation of available funds toward hiring employees, covering day-to-day operational expenses, or managing inventories. Consequently, a significant portion of the funds may be utilized for debt repayment or operational expenses rather than expanding the asset base.

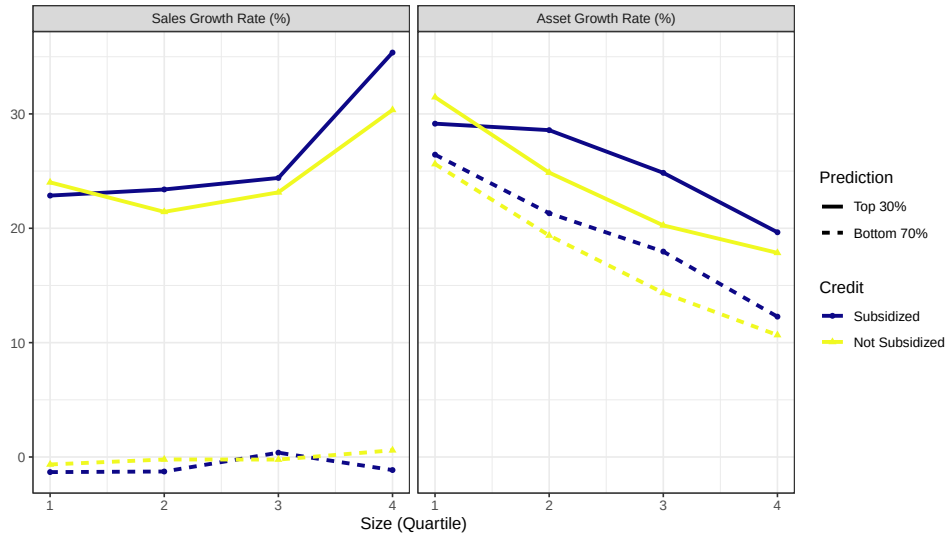

Figure 7: Firm performance comparison by ML prediction and debt-to-capital ratio
Notes: The figure displays the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target and quantiles of debt-to-capital ratio defined at each year and industry group. The debt-to-capital ratio is defined at the firm level as the ratio of total liabilities to total capital.

However, there are concerns regarding the use of debt-to-capital ratio and cash flow ratio as criteria for credit allocation. These measures are influenced by endogenous financial choices made by the firms themselves. For instance, a high debt-to-capital ratio may simply indicate the firm's ability to access debt financing rather than being a clear indicator of credit constraints. Furthermore, these measures are more susceptible to manipulation compared to firm size and age. For example, a firm with the capability to borrow from external markets may intentionally increase its debt-to-capital ratio to become eligible for subsidized loans. Similarly, although it is not common, firms could potentially lower their operating income

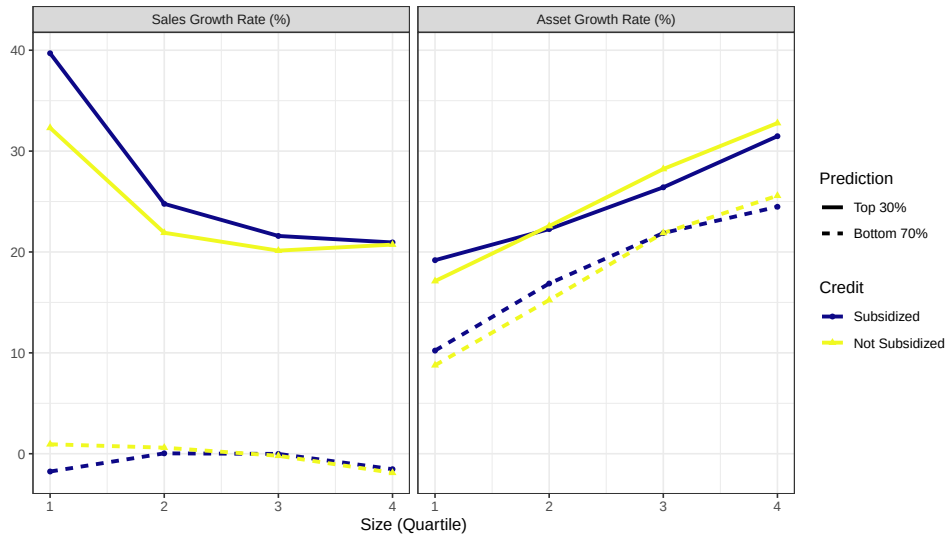

Figure 8: Firm performance comparison by ML prediction and cash flow ratio
Notes: The figure displays the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target and quantiles of cash flow ratio defined at each year and industry group. Cash flow ratio is operating income over lagged tangible assets.

by deliberately inflating expenses or writing off inventory in order to lower their cash flow ratio.

Considering these concerns, we find that the firm's size and age are particularly useful indicators to be utilized for credit allocation. These factors provide more objective and reliable information that is less prone to manipulation. By incorporating firm size, age, and sales growth predictions into the decision-making process, policymakers can achieve more effective resource allocation in support of credit-constrained firms.

6 Conclusion

This paper asserts that the effectiveness of government-subsidized credit programs for SMEs can be significantly enhanced through a data-driven approach by targeting credit-constrained but viable firms. Based on an administrative database that records all subsidies to SMEs, we

propose a machine learning model that predicts annual sales growth rates to identify the most appropriate beneficiaries. Our study contributes to the growing studies that demonstrated the efficacy of data-driven decision-making in public policies. Our research highlights the potential of machine learning applications in resource allocation problems for various business support policies.

When we looked at the actual growth rates of sales, the bottom 70% predicted sales growth group exhibited negligible average sales growth, and the lowest 30% predicted group experienced negative growth, indicating the allocation of resources to zombie firms. These results highlight the potential of ML predictions to increase the effectiveness and cost-efficiency of government-subsidized credit programs. Moreover, our study offers empirical evidence supporting the use of firm size and age as reliable indicators of credit constraints, which can further improve allocation efficiency when combined with predicted sales growth.

Our approach can be widely applicable to other public business support programs to identify their suitable targets. The future growth prediction can be used as a supplementary or primary criterion to screen the candidates in the screening stage. We demonstrated that an ML model can be applied to new applicant firms based on their observable information, as the model is estimated using currently available data to policymakers. Although the Korean government has introduced a data integration platform to manage data on SME support programs, it has not been effectively used to improve program efficiency. We provide a case study of data-driven decision-making where we utilize integrated administrative data combined with firm-level data. Nonetheless, it is important to recognize the limitations of ML algorithms and potential issues that may arise during implementation.

To enhance transparency during policy implementation, it is crucial to address related issues. First, our findings could be susceptible to omitted-variable biases if variables correlating with firms' growth were omitted.[12] The credit program managers might have considered

[12] Kleinberg, Lakkaraju, Leskovec, Ludwig and Mullainathan (2018) emphasized that understanding the

additional objectives other than just the viability of firms, such as saving them from default. We found that the sales growth rates of firms with low predicted growth were even lower for the subsidized group than not subsidized group. Human decisions that seemed unfavorable to the growth of firms could have been made to save them from going bankrupt and the resulting loss of jobs. However, given the primary objectives of government credit support policies, these programs must target firms that exhibit strong growth potential once they receive credit. By doing so, policymakers can evaluate the programs' effectiveness in achieving their goals. Additionally, the Korean government operates business rescue programs that help financially distressed businesses recover and avoid bankruptcy. Since these rescue programs provide credit support to those struggling firms, policymakers can minimize inefficiencies caused by omitted-variable biases by explicitly distinguishing their objectives for supporting businesses.

Secondly, our model is limited in its ability to assist policymakers in allocating credits between industries. While our model is estimated for each industry and year and predicts firm's future performance relative to other firms within the same industry, it does not provide guidance for credit allocation across different industries. When policymakers assign equal importance to all industries, credits can be allocated based on the prediction results. However, policymakers can consider the size and impact of industries on the economy to inform credit allocation decisions. Therefore, an additional model or approach may be required to address the allocation problem between industries.

Third, in order to design an effective decision aid based on the ML prediction algorithm, a trial process and evaluation are needed. It may be difficult to strictly rely on the ML prediction for eligibility criteria, as there can be large variations in sales growth rates. Moreover, the interpretation of the decision made by the model can be challenging, as it works like a black box. Therefore, further research could explore ways to enhance model interpretability

omitted-payoff biases is important to improve decision quality based on prediction.

to provide outcomes that are readily understandable.

One possible approach is to introduce a two-step verification process for firms whose sales growth rates are predicted to be in the bottom 30%. Our findings showed that their sales declined after receiving credits, indicating that additional verification for their creditworthiness is necessary to identify firms that are suitable for policy support. By undergoing a trial and post-evaluation process, the ML algorithm can be a more effective decision aid in improving policy effectiveness.

References

Andini, Monica, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, and Viola Salvestrini, "Targeting with machine learning: An application to a tax rebate program in Italy," *Journal of Economic Behavior & Organization*, 2018, *156*, 86–102.

_ , Michela Boldrini, Emanuele Ciani, Guido De Blasio, Alessio D'Ignazio, and Andrea Paladini, "Machine learning in the service of policy targeting: the case of public credit guarantees," *Journal of Economic Behavior & Organization*, 2022, *198*, 434–475.

Arata, Yoshiyuki, "Firm growth and Laplace distribution: The importance of large jumps," *Journal of Economic Dynamics and Control*, 2019, *103*, 63–82.

Bach, Laurent, "Are small businesses worthy of financial aid? Evidence from a French targeted credit program," *Review of Finance*, 2014, *18* (3), 877–919.

Banerjee, Abhijit V and Esther Duflo, "Do firms want to borrow more? Testing credit constraints using a directed lending program," *Review of Economic Studies*, 2014, *81* (2), 572–607.

Berryhill, Jamie, Kévin Kok Heang, Rob Clogher, and Keegan McBride, "Hello, World: Artificial intelligence and its use in the public sector," *OECD Working Papers on Public Governance*, 2019, (36).

Bertoni, Fabio, Jose Martí, and Carmelo Reverte, "The impact of government-supported participative loans on the growth of entrepreneurial ventures," *Research Policy*, 2019, *48* (1), 371–384.

Blasio, Guido De, Stefania De Mitri, Alessio D'Ignazio, Paolo Finaldi Russo, and Lavinia Stoppani, "Public guarantees to SME borrowing. A RDD evaluation," *Journal of Banking & Finance*, 2018, *96*, 73–86.

Bottazzi, Giulio and Angelo Secchi, "Explaining the distribution of firm growth rates," *The RAND Journal of Economics*, 2006, *37* (2), 235–256.

Brown, J David and John S Earle, "Finance and growth at the firm level: Evidence from SBA loans," *The Journal of Finance*, 2017, *72* (3), 1039–1080.

Clementi, Gian Luca and Hugo A Hopenhayn, "A theory of financing constraints and firm dynamics," *The Quarterly Journal of Economics*, 2006, *121* (1), 229–265.

Cowan, Kevin, Alejandro Drexler, and Álvaro Yañez, "The effect of credit guarantees on credit availability and delinquency rates," *Journal of Banking & Finance*, 2015, *59*, 98–110.

Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda, "The role of entrepreneurship in US job creation and economic dynamism," *Journal of Economic Perspectives*, 2014, *28* (3), 3–24.

Fairlie, Robert W, Dean Karlan, and Jonathan Zinman, "Behind the GATE experiment: Evidence on effects of and rationales for subsidized entrepreneurship training," *American Economic Journal: Economic Policy*, 2015, *7* (2), 125–161.

Green, Anke, *Credit guarantee schemes for small enterprises: an effective instrument to promote private sector-led growth?*, UNIDO, Programme Development and Technical Cooperation Division, 2003.

Hadlock, Charles J and Joshua R Pierce, "New evidence on measuring financial constraints: Moving beyond the KZ index," *The Review of Financial Studies*, 2010, *23* (5), 1909–1940.

Haltiwanger, John, Ron S Jarmin, and Javier Miranda, "Who creates jobs? Small versus large versus young," *Review of Economics and Statistics*, 2013, *95* (2), 347–361.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Vol. 2, Springer, 2009.

Hottenrott, Hanna and Robert Richstein, "Start-up subsidies: Does the policy instrument matter?," *Research Policy*, 2020, *49* (1), 103888.

Hu, Yunzhi and Felipe Varas, "A theory of zombie lending," *The Journal of Finance*, 2021, *76* (4), 1813–1867.

Huergo, Elena and Lourdes Moreno, "Subsidies or loans? Evaluating the impact of R&D support programmes," *Research Policy*, 2017, *46* (7), 1198–1214.

Hurst, Eric and Ben Pugsley, "What Do Small Businesses Do?," *Brookings Papers on Economic Activity*, 2012.

Kaplan, Steven N and Luigi Zingales, "Do investment-cash flow sensitivities provide useful measures of financing constraints?," *The Quarterly Journal of Economics*, 1997, *112* (1), 169–215.

Kim, Minho, "Aggregate productivity growth in Korean manufacturing: the role of young plants," *KDI Journal of Economic Policy*, 2017, *39* (4), 1–23.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

_ , Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, "Prediction policy problems," *American Economic Review*, 2015, *105* (5), 491–495.

Kwon, Hyeog Ug, Futoshi Narita, and Machiko Narita, "Resource reallocation and zombie lending in Japan in the 1990s," *Review of Economic Dynamics*, 2015, *18* (4), 709–732.

Lagazio, Corrado, Luca Persico, and Francesca Querci, "Public guarantees to SME lending: Do broader eligibility criteria pay off?," *Journal of Banking & Finance*, 2021, *133*, 106287.

Lamont, Owen, Christopher Polk, and Jesús Saaá-Requejo, "Financial constraints and stock returns," *The Review of Financial Studies*, 2001, *14* (2), 529–554.

Liaw, Andy and Matthew Wiener, "Classification and Regression by randomForest," *R News*, 2002, *2* (3), 18–22.

McKenzie, David and Dario Sansone, "Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria," *Journal of Development Economics*, 2019, *141*, 102369.

Ministry of SMEs and Startups, *Guidebook for 2018 SMEs and Venture Business Support Programs (in Korean)* 2018.

Mullainathan, Sendhil and Jann Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

National Assembly of the Republic of Korea, "National Finance Act," National Assembly of the Republic of Korea 2023. Article 85(7).

Organisation for Economic Co-operation and Development, *Artificial Intelligence in Society* 2019.

_ , *Financing SMEs and Entrepreneurs 2020* 2020.

R Core Team, *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing 2022.

Sansone, Dario and Anna Zhu, "Using Machine Learning to Create an Early Warning System for Welfare Recipients," *IZA Discussion Paper No*

14377, 2021. available at https://www.iza.org/publications/dp/14377/using-machine-learning-to-create-an-early-warning-system-for-welfare-recipients.

Small and Medium Business Administration, *Fiscal Year 2015 Summary of Revenue and Expenditures, and Programs Report (in Korean)* 2016.

Ubaldi, Barbara, Enzo Maria Le Fevre, Elisa Petrucci, Pietro Marchionni, Claudio Biancalana, Nanni Hiltunen, Daniela Maria Intravaia, and Chan Yang, "State of the art in the use of emerging technologies in the public sector," *OECD Working Papers on Public Governance*, 2019, (31).

Varian, Hal R, "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, 2014, *28* (2), 3–28.

Whited, Toni M and Guojun Wu, "Financial constraints risk," *The Review of Financial Studies*, 2006, *19* (2), 531–559.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani, "Welcome to the tidyverse," *Journal of Open Source Software*, 2019, *4* (43), 1686.

Zia, Bilal H, "Export incentives, financial constraints, and the (mis) allocation of credit: Micro-level evidence from subsidized export loans," *Journal of Financial Economics*, 2008, *87* (2), 498–527.

APPENDIX

A Appendix Tables and Figures

Table A.I: Annual firm performance comparison by ML prediction

Loan status	ML predicted growth	Year	Sales growth rate(%)	Asset growth rate(%)	*N*
Subsidized	Top 30%	2011	28.2	22.5	2715
		2012	28.2	24.1	2448
		2013	27.7	26.5	2460
		2014	28.8	24.3	2106
		2015	26.3	25.5	2035
	Bottom 70%	2011	-1.7	16.6	5793
		2012	-0.4	20.9	4136
		2013	-1.6	19.9	3764
		2014	-0.4	21.0	2746
		2015	1.3	23.5	2381
Not Subsidized	Top 30%	2011	30.9	26.1	4586
		2012	30.0	27.2	4699
		2013	26.3	24.9	4755
		2014	20.4	22.1	4584
		2015	17.4	14.8	3609
	Bottom 70%	2011	1.1	21.2	10462
		2012	1.2	19.7	11482
		2013	-1.0	19.1	12114
		2014	-1.3	17.3	12119
		2015	-0.7	14.3	10700

Note: The table reports the average growth rates of sales and total assets for both subsidized and not subsidized groups of firms by ML predicted target and year. N is the number of observations.