Aim Better with Machine Learning: Enhancing Effectiveness of Government Credit Programs for SMEs*

Minho Kim[†] Korea Development Institute

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

Abstract

Government-subsidized loans and public credit guarantees are key policy measures that provide financial support to small and medium-sized enterprises (SMEs). The effectiveness of these policies hinges on properly targeting those aligned with the policy goal of supporting firm growth. However, current government credit programs are often undermined by adverse selection and insufficient screening of recipient firms due to information asymmetries and moral hazards. This paper utilizes South Korea's administrative database of government SME credit support programs together with observable firm characteristics to build a prediction model for effectively identifying high-growth firms. We demonstrate that such a data-driven approach can potentially improve the effectiveness of government credit programs and can be easily applied to other grants or funds to support business growth. Additionally, our findings reveal that firm size and age serve as highly informative indicators of credit constraints.

*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

JEL Codes: G21, G28, H81, L26

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

1 Introduction

Government-subsidized loans and public credit guarantee programs are employed as cornerstone industrial policies in many countries, aiming to support financially challenged small and medium-sized enterprises (SMEs). These programs act as a catalyst for SMEs' growth by easing their access to capital. This is crucial because SMEs, especially startups and small enterprises, often face difficulties securing funding due to limited ability to demonstrate their creditworthiness, leading to higher interest rates or loan denials. By intervening in the credit market, government financing mitigates information asymmetries that hinder lending to credit-constrained SMEs, ultimately facilitating their business expansion.

However, despite aiming to support promising firms with temporary credit constraints, these programs can attract a disproportionate number of "lemons" – beneficiaries who do not genuinely need the support.[1] This issue arises due to two factors. First, 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. Second, banks, knowing that public credit programs guarantee most of the debts (e.g., 85 - 100% in South Korea), may prioritize riskier borrowers and engage in adverse selection with less rigorous screening and monitoring (moral hazard, Cowan et al., 2015; De Blasio et al., 2018; Lagazio et al., 2021).

These challenges from both borrowers and lenders often lead to a sub-optimal selection of

[1]The lemons problem fundamentally arises from asymmetric information between lenders and borrowers. The issue of firms' incentive problems, under conditions of asymmetric information, related to government credit programs has been explored in early studies, including those by (Chaney and Thakor, 1985; Innes, 1991; Gale, 1990, 1991).

recipient firms, effectively adding a burden to public finance and falling short of the intended outcomes. The fundamental challenge is thus clear: How can we effectively identify those firms experiencing credit constraints while possessing high growth potential once supported by government support?

This study demonstrates the potential of a data driven approach to significantly improve the effectiveness of government credit support programs, which is achieved by screening out unsuitable SMEs using a simple machine learning model. Utilizing the model prediction of the sales growth of recipient firms, such an approach can address aforementioned two main problems: adverse selection and poor screening. Leveraging an extensive dataset on policies aimed at young firms, we show that allocating credits based on predicted sales growth can be an effective and straightforward method for ruling out less viable applicants, leading to a more efficient and impactful allocation of public resources.

We combine four large-scale administrative databases covering nationwide government-guaranteed and direct government loans provided between 2010 and 2015 with annual finance and basic characteristics data of individual firms. This integration creates a longitudinal firm level dataset with 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 the ongoing policy shift, where numerous governments are moving towards data-driven decision-making in public services, often advocating for incorporating Artificial Intelligence (AI, Berryhill et al., 2019). This shift aims to personalize government services and improve their effectiveness. As large-scale datasets become more widely available, 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). Many studies examined 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; Horvath and Lang, 2021). While causal analysis remains valuable as it allows us to examine the impact of policy interventions, our study explores the possibility of policy improvement by leveraging the predictive outcomes from the ML model. Employing predictive models aids in preventing the misallocation of public resources to failing firms ("zombie lending," Kwon et al., 2015; Caballero et al., 2008; Hu and Varas, 2021) and instead directs them towards high-potential firms that can can significantly contribute 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 can capture complex interactions and nonlinear structures (Varian, 2014). This characteristic of ML has led to a growing body of research beyond engineering perspectives, so that its application in resource allocation decision-making can be considered 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 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%). Focusing on the bottom 30% of firms categorized by the predicted growth rate, we observed a notable decrease 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, our analysis reveals that firm size and age emerge as notably effective indicators for identifying credit constraints. This paper contributes to the scholarly discourse surrounding firm growth, the implications of credit constraints, and the role of 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 distinguishes itself from prior studies by focusing on the incorporation of firms' credit constraints into real-world credit allocation processes. We explored an array of metrics for assessing financial constraints, as proposed in the corporate finance literature.

While our study highlights the benefits of ML predictions in policy making, it's crucial to acknowledge their limitations and exercise caution (McKenzie and Sansone, 2019). To

ensure transparency and reliability for policy implementation, we examine the limitations of ML tools. Recognizing these limitations can help the policymakers to address potential shortcomings of these tools, 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

Numerous governments have implemented policy instruments to facilitate SMEs' access to financing, thereby enabling their growth and job creation. Since the 19th century in Europe, credit guarantee schemes have been widely adopted by both developed and developing countries to support private businesses. These schemes are operational in nearly 100 countries (Green, 2003). Additionally, 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(SBA)

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 are designed to support the growth of companies possessing strong technological capabilities and significant business potential, yet are constrained by limited financial resources.[2] Public guarantee schemes function to mitigate the difficulties faced by SMEs in securing credit from the market. They achieve this by shifting the risk associated with lending from financial institutions to public entities. 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 expanded credit guarantee schemes or direct government loans to support SMEs.[3]

In Korea, 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.[4] The interest rate for subsidized loans has consistently remained between 1% and 2.2% lower than the market rate, as observed in a quarter-wise comparison from the fourth quarter of 2012 to the fourth quarter of 2017.[5] 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.

In our case study, the South Korean government's credit programs include both loan

[2]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).

[3]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).

[4]Relative to the SBA's principal business loan initiative, the 7(a) Loan Program in the United States, the terms concerning the guarantee percentage of loans in Korea are comparatively more favorable. Specifically, for the majority of loans under the 7(a) program, the SBA guarantees up to 85% for loans amounting to $150,000 or less. For loans exceeding $150,000, the guarantee is up to 75%.

[5]The rates for subsidized loans are variable and determined every quarter, being tied to the policy fund's base rate. The policy fund's base rate is, in turn, linked to the SME promotion bond procurement rate.

guarantees and subsidized loans, explicitly targeting SMEs that (1) exhibit high growth potential but (2) face significant constraints in accessing finance for their investments. However, evaluating an applicant's growth potential poses a considerable challenge, while determining the extent of credit constraints is even more intricate. Generally, within the credit application procedure, the assessment of an applicant's growth potential relies on a qualitative evaluation. This evaluation encompasses a review of the feasibility of their technological and business plans, along with their management skills. The programs do not take into account the potential availability of loans for a candidate firm from private banks during their decision-making process. However, it is explicitly stipulated that certain large enterprises, such as publicly listed companies or those with annual sales exceeding 50 billion KRW in the preceding financial year, are ineligible.

Firms may seek government credit support to benefit from preferential interest rates, even in the absence of credit constraints. Another reason may be to sustain their operations when they fail to innovate or adapt to market changes. Coupled with lenient screening processes, these workarounds can result in substantial inefficiencies in government credit programs. Rather than being allocated to productive firms facing genuine credit constraints, government credits may be tied up in zombie firms characterized by stagnant or diminishing revenue streams.

The challenge of identifying the appropriate policy target can be effectively mitigated through the application of ML techniques, which use sales growth as the target outcome variable (dependent variable) for prediction purposes. This method transcends the limitations of purely qualitative assessments. By employing a machine learning-based approach, it is possible to more precisely identify firms that are credit-constrained yet possess potential for growth, as indicated by the predicted rate of sales growth.

Furthermore, moving beyond the passive strategy of merely excluding large corporations from policy targets, our approach involves identifying and utilizing robust indicators to

gauge the extent of financial constraints faced by businesses. This methodology enables a more strategic allocation of credits, ensuring a more efficient and impactful distribution of financial support, aligning with the overarching objectives of the credit programs.

3 Selecting Suitable Targets for Government Credit Programs

In this section, we develop the theoretical rationale and illustrate through 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 are 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 in terms of the logarithmic increase in sales.

Banerjee and Duflo (2014) shows that "if a 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."[6] 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 allocated to firms that are not credit-constrained, their 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 an elevated growth rate when they receive 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 degree of credit constraint and the 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 by the overarching objectives of the policy. The ratio $\frac{\alpha_w}{\alpha_y}$ is larger when the policy prioritizes the allocation of credit to firms experiencing more significant credit constraints, as opposed to those firms exhibiting higher growth potentials. The policy decision is to establish eligibility criteria based on the firm's potential for 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 $\int_{\Omega_c} z dP_{w,y|y\in\mathcal{S}_c} > \int_{\Omega_{c'}} z dP_{w,y|y\in\mathcal{S}_{c'}}$.*

[6] 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.

Note that $\pi > 0$ implies that, at the decision-making stage, it has been predetermined 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 $\sigma_*^2 = (1 - \rho^2)\sigma_W^2$. 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 illustrates a hypothetical distribution of applicant firms, wherein each dot symbolizes a potential applicant. These dots are plotted based on two key variables: the level of credit constraint ($cc_{i,t}$) and the predicted growth ($y_{i,t}$). The level of credit constraints for each firm is represented through the color density of the dots, where a darker hue indicates a lower level of credit constraint. In particular, we assumed that the potential growth rates

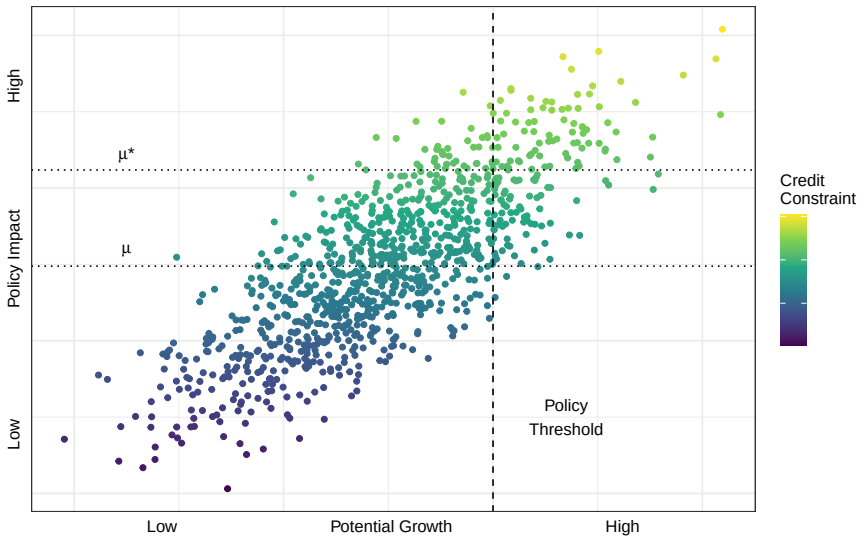

Figure 1: A Simulated Example: Policy Impact on Firms Distributed by Predicted Growth and Credit Constraints

and the level of credit constraints are jointly normally distributed, with these two variables exhibiting a positive correlation. [7]

The findings presented in Proposition 1 indicate that, under certain specified conditions, the expected policy impact increases when an eligibility criterion is implemented. This criterion extends credit availability to firms projected to surpass a pre-established growth threshold. As depicted in Figure 1, the policy threshold is represented by a solid line. Credit is allocated to firms whose expected growth is situated to the east of this threshold line. Comparatively, the expected policy impact when employing a threshold, denoted as $\mu_(c^*)$, surpasses the expected policy impact in the absence of such a threshold, represented by 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. Government credit programs typically operate within the confines of annual budgetary constraints. Within these

[7]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). It can be shown that the mechanism illustrated here is still applied for Laplace distributions with a slight modification.

fiscal boundaries, various approaches can be employed to determine the eligibility of firms for credit allocation.

One approach involves ranking firms according to their projected growth rates, enabling the program to allocate funds to those applicants ranked highest within the budget. Alternatively, the program can establish a specific growth rate threshold, such as zero or the industry's average growth rate, as a criterion for eligibility. Acknowledging the inherent uncertainties in growth predictions, the program could further introduce a supplementary evaluation process. This would allow firms falling just below the threshold an opportunity for reconsideration, thereby providing a safeguard against premature exclusion based solely on initial projections.

An additional rationale for targeting the sales growth rate lies in its role as a key performance indicator (KPI). This metric is instrumental in assessing the efficacy of government support initiatives. The Korean Ministry of SMEs and Startups has explicitly identified sales growth rates as the primary performance indicators in its annual Performance and Accountability Report.[8] This report comprehensively outlines the program's objectives alongside the corresponding performance indicators. Following this, it presents an assessment of actual performance, which plays a vital role in financial management and evaluating the efficacy of policies.[9] By targeting the sales growth rate, a 'what-if' analysis can be conducted to gauge the policy's impact. This approach offers policymakers a valuable tool to discern the most suitable targets from the applicant pool.

[8]Cite: See Small and Administration (2016). 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).

[9]The growth in total assets can also be considered a predicted variable given that loans are 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.

4 Data and modeling methodology

4.1 Data

Our model is developed upon two data sources. The first source comes from the Integrated Management System for Small and Medium Enterprise Aid Programs (SIMS) database. This administrative database provides comprehensive yearly information on subsidies given to SMEs by any central and local governments. The SIMS database records data on project names, functional areas, departments accountable, budgets, and corresponding subsidy amounts at the firm level from the year 2010.[10] However, this database lacks important details such as balance sheet data and basic characteristics of the SMEs. To address this limitation, we integrated the SIMS data with information from the Korea Enterprise Data (KED) database, renowned as South Korea's largest repository for SME credit information.[11] The KED database provides information on revenue, assets, profits, capital, liability, industry, establishment year, and corporate forms of business ownership. We link the annual firm-level data from the SIMS with the KED database using the business registration number.

Our study focuses on the firms that received government-guaranteed loans and direct government loans from 2010 to 2015. Four national programs share a similar objective: to provide financing to SMEs demonstrating potential for growth. 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. During the period from 2010 to 2015, these four programs collectively

[10] Korean government introduced the SIMS to improve the efficiency of SME support policies.
[11] see http://www.kodata.co.kr/en/ENINT01R1.do

provided guarantees or loans to over 80,000 firms annually. By utilizing the SIMS-KED integrated data, we identify firms that received loans from any of these four major programs and treat them as a subsidized group for our study.

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, this analysis can partially account for the varying growth patterns observable across different firm age groups. Empirical research has demonstrated that younger firms tend to exhibit substantially higher mean growth rates and greater dispersion in these rates compared to their older counterparts, conditional upon 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 for 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.

Utilizing this dataset, we construct separate datasets for each combination of three-digit industry and year, as we estimate a ML model for each pair. Although there is no universally established minimum data size for ML models, we exclude any pair comprising fewer than 500 firms. This criterion is to ensure that each model has sufficiently large data to discern patterns from the covariates. Companies lacking balance sheet information from the preceding year are excluded from our analysis. In alignment with the size eligibility

criteria for government loans to SMEs, we excluded companies whose sales exceeded KRW 50 billion or whose total assets surpassed KRW 100 billion in the preceding year. Under other eligibility criteria, companies listed on the Korea Exchange (KRX) or the KOSDAQ market, along with cooperatives, and non-profit organizations are also excluded.

After applying the aforementioned criteria, the data was narrowed down to 395,541 firm observations from 2010 to 2015. During this period, approximately 17.6% of the firms received their first government-subsidized loan. We assert that the data is apt for analysis, as it encompasses the majority of the SMEs in Korea and includes extensive records of government credit support provided to these SMEs.

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

On average, the sales and profits of subsidized firms align closely with those of 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, implying operation on relatively constrained capital, potentially due to financial constraints. Moreover, these firms exhibit a higher debt-to-capital ratio and lower cash flow, signaling potentially greater credit constraints. The combination of their younger age and smaller asset size underscores a heightened susceptibility to 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. In the following section, we describe the process of utilizing a machine learning (ML) model on our dataset, which includes the deployment of a matching algorithm. This technique is

specifically employed to select a subset of non-subsidized firms that exhibit characteristics comparable to those of the subsidized firms.

Table 1: Summary Statistics

	Subsidized		Not Subsidized	
Variable	Mean	SD	Mean	SD
Sales (1M)	3399	(8189)	3295	(7185)
Operating income (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. Markups are equal to sales divided by costs (sales minus operating income).

4.2 Predictive model

ML models are particularly effective in learning relationships between a target variable and numerous predictor variables, which may often be nonlinear, due to their ability to accommodate flexible functional forms and to account for interaction effects (Varian, 2014). As implied by its name, a ML model can automatically 'learn' such relationships, which becomes particularly advantageous in cases marked by high dimensionality with many predictor variables. ML models can be designed to adapt to new data, offering a significant advantage in dynamic environments including annual policy target selection.

Our modeling approach aims to *predict* the performance of a firm in the future, given the

information available at the current year. If there are critical factors that affect the firm's future growth, the model should be able to reflect this underlying structure by incorporating the impact of these factors. Our approach directly addresses the problem from a decision-making standpoint: making future predictions with the information available at the time of selecting the subsidy beneficiaries.

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 structures (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 a 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) 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 2023 to predict 2024's performance, potential data contamination can be minimized.

Third, the initial data set is built as an annual data set consisting of sales, operating income, total assets, tangible assets, liability, paid-in capital, venture firm status, industry, number of full-time employees, age, subsidized loan status, and corporate forms of business ownership for every year.[12] The data set is re-structured so that for every annual sales growth, the input variables contain information on firms up to two years before their application. The input variables are listed in the appendix A.I. Firm-level markups are included to capture market or regulatory distortions (see Baqaee and Farhi (2020)). Variables potentially indicative of credit constraints, such as the cash flow ratio and debt-to-capital ratio, are also included.

Lastly, we conduct a screening within each industry block to select a subset of firms closely

[12]In the initial data wrangling process, other variables that have more than 20% missing were omitted from the analysis. Business ownership is categorized into several types: listed on the 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.

resembling the subsidized firms in terms of relevant characteristics. By applying the Nearest Neighbor Matching method, we identify the five nearest unsubsidized neighbors in the input space for each subsidized firm. The dataset for each block is then formed by the union of these selected firms and the subsidized ones. This matching process, while reducing the total number of firms in the modeling and validation stages, presents a challenge by reducing the sample size and increasing the difficulty in distinguishing between firms. Nonetheless, it ensures that the prediction model is estimated for firms sharing characteristics similar to the applicants.

When presenting our results, we include the outcomes of non-subsidized firms for comparison purposes. Utilizing matching techniques facilitates more fair comparison by omitting firms whose characteristics are markedly distant from those of the subsidized firms.[13]

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

[13]By using the matching procedure, we try to address the issue of selection bias and endogeneity that may arise due to the correlation between the characteristics of credit applicants and their actual outcomes. However, we do not view this problem as a significant concern, as our prediction model is exclusively applied to the applicants.

– the 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 not be very accurate ('weak'). RF builds many weak trees, makes a prediction for the new target data from each tree, and averages them to make the final prediction. Variance of the final prediction tends to be reduced by aggregating predictions from multiple models. This reduction is due to weak correlations among different trees because each tree uses distinct 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

The model's performance is assessed by its prediction accuracy for firms' sales growth. [14] 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 by predicted growth rate show an average annual sales growth rate of approximately 50%, compared to around 19% for the next 10%. On the contrary, firms projected to fall within the three lowest groups show negative growth rates on average. Moreover, subsidized

[14]The appendix A.I lists the average rankings of variables for each industry-year block by their variable importance. This arrangement shows their relative contributions to the model's prediction.

firms predicted to be in the top performing group showed marginally higher growth rates compared to non-subsidized firms. The findings demonstrate the efficacy of ML in identifying firms that are likely to perform well in the following year.

It is worth noting that we base our identification of high growth firms on their predicted values relative to others within the same industry. Empirical research has highlighted the challenges of predicting firm growth due to significant variations in annual growth rates observed within firms over time (Coad and Srhoj, 2020). When evaluating the prediction accuracy at the individual firm level, the squared correlation (R^2) between the predicted and actual outcome is 9.0%.[15] By focusing on the average outcomes based on the relative predicted performance, our findings indicate that predictive outcomes furnish decision-makers with valuable information to improve policy effectiveness. Specifically, we offer a tool to exclude firms exhibiting 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 to 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 rare for some firms to experience negative sales growth in a given year, providing subsidies to firms projected to decline would be a hard sell to taxpayers. Predictive

[15]The squared correlation obtained from the linear regression, employing the same covariates as those used in the ML model, was 2.6%. This magnitude bears resemblance to the findings of Coad and Srhoj (2020), where the LASSO (Least Absolute Shrinkage and Selection Operator) technique was utilized to predict high growth firms.

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 instance, they can reassess the provision of loans to firms that fall within the lowest 30% of the predicted performance group. Alternatively, loans can be reallocated based on predicted performance groups, ensuring that firms in the lower groups receive less funding overall. In Section 5 that follows, we explore the quantitative benefits of utilizing ML predictions in a government loan program.

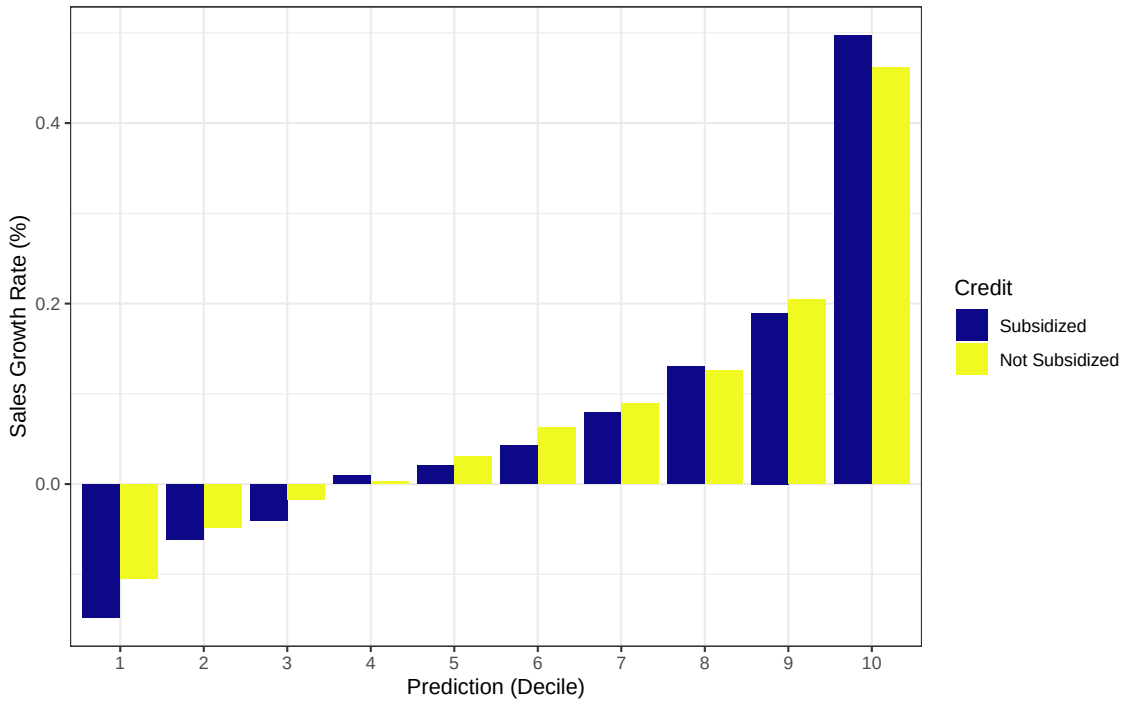

Figure 2: Machine Learning Predictions and Actual Outcomes
Notes: The figure shows the average growth rate of sales for both subsidized firms and not subsidized firms across deciles of predicted growth.

5 Effectiveness of ML predictions and human decision

In this section, we explore the application of the predicted information in the actual policy-making stage to select the loan recipients. We suggest the adoption of a simple yet effective decision criterion: grant a loan to a firm only if its predicted sales growth rate is within the highest 30% of all the applicants. It is important to note that the 30% threshold serves as an illustrative example. The actual threshold can be determined by considering various factors, such as the amount of subsidy available and the growth objectives that the policy aims to achieve. Beyond the common practice in support projects of using recipients' sales growth rate as a performance indicator, this measure is preferred as it is reasonable to expect that firms will expand their sales when they are credit-constrained yet possess growth prospects. Moreover, sales is a direct performance indicator that has less room for a firm to manipulate compared to other measures such as operating income. In the following section 5.1, we apply the simple allocation rule providing loans to the top 30% of applicants based on predicted sales growth.

Another important dimension in the selection of loan recipients is the extent of their credit constraints. 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 production or investment after they substitute subsidized credit for market credit because subsidized credit is cheaper than market credit (Banerjee and Duflo, 2014). Sales are likely to 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 examine various proxy variables indicative of credit constrainedness, combined with the predicted sales growth rate.

5.1 Utilizing predictions from machine learning to find targets

We conduct a quantitative assessment of the impact that machine learning has on improving the efficiency of subsidized loan allocation. We identify potential recipients using the ML predictive model, as detailed 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. The appendix A.II 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 compares the results for firms that received subsidized loans and those that did not get any subsidized 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.

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. Although we have chosen non-subsidized firms with characteristics similar to those of subsidized firms, there may still

be unobserved characteristics that differ between these groups. They might be more credit-constrained or have a higher incentive to substitute their existing credit with subsidized credit. Thus, we refrain from inferring any causal effects of receiving subsidized loans on 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.

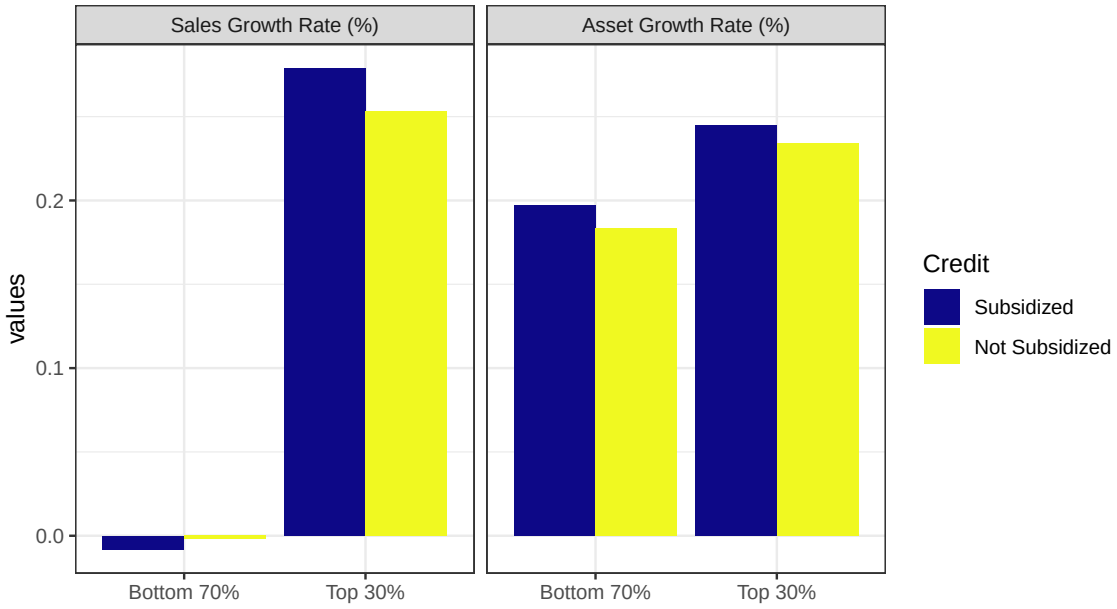

Figure 3: Firm performance comparison by ML prediction target
Notes: Panels (a) and (b) display the average growth rates of sales and total assets, respectively, for firms in the top 30% and bottom 70% based on the predicted sales growth. The average values are presented for both subsidized and non-subsidized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

5.2 Considering credit constrained firms

Once we replace the subjective assessment of the applicant firms with their predicted sales growth, we can explore ways to consider credit-constrainedness of these firms when selecting loan recipients. Our study employs four measures, as suggested in the corporate finance liter-

ature, to assess the financial constraints of the firms, which are firm size, age, debt-to-capital ratio, and cash flow ratio. This analysis has the potential to enhance the efficiency 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 credit support.

Various methods were proposed to measure the extent of a firm's credit constraints. 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) developed an index of a firm's finance constraints using the generalized method of moments (GMM) for the estimation of a structural model. These studies utilized variables extracted from accounting information to estimate the degree of financial constraints. These variables included operating income or cash flow over tangible assets, Tobin's Q, debt-to-capital ratio, dividends, asset size, and firm or industry sales growth rate.

However, when Hadlock and Pierce (2010) re-examined the issue of measuring credit constraints with newly available data, they identified inconsistencies in previously used variables. 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, particularly the leverage ratio, were subject to limitations due to endogeneity issues, potentially leading to biased estimations. Consequently, they proposed that a firm's asset size and age serve as useful proxies for assessing the degree of financial constraints since they are consistent and less susceptible to endogeneity problems.

Following 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 examine a firm's debt-to-capital and cash flow ratios, as demonstrated by Hadlock and Pierce (2010) for their consistency in predicting a firm's constraint status. For each year and industry group, firms are categorized into quartiles based on three of the four measures of constraints, while age is used in its original form. Subsequently, we compare the average actual growth rates of sales and assets across these categories to assess whether more credit-constrained firms exhibit higher growth rates.

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. In the high-growth expected group, subsidized firms exhibited higher sales growth rates than non-subsidized firms across all firm size quartiles. 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

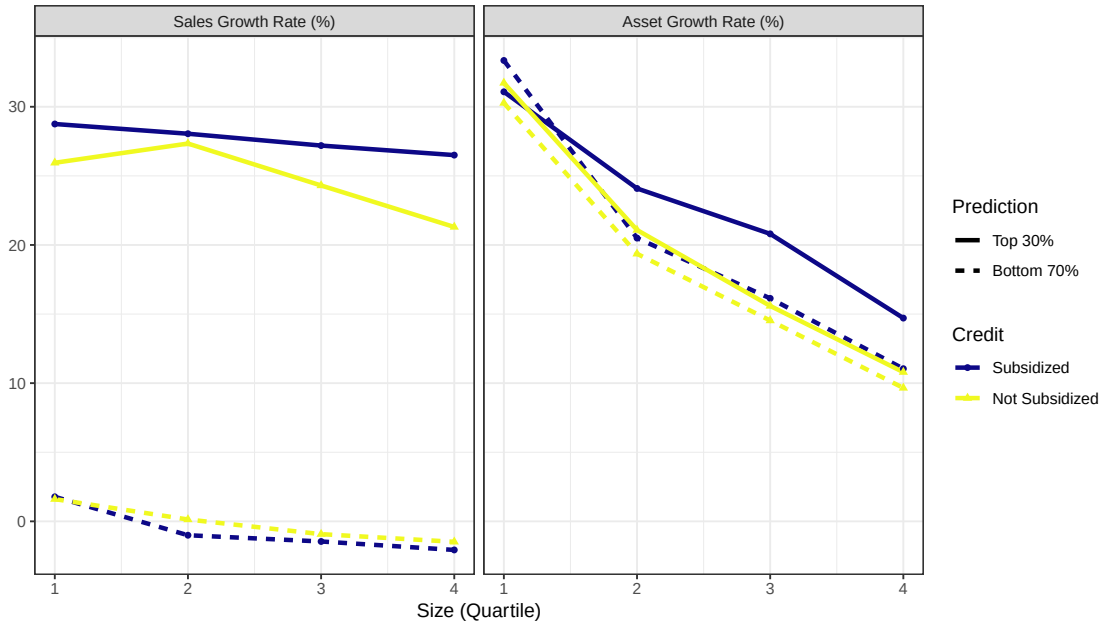

Figure 4: Firm performance comparison by ML prediction and size
Notes: The figure displays the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target and asset size quantiles defined at each year and industry group.

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 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. The sales growth rate for subsidized firms was generally lower than that for not-subsidized firms across most age groups.

Remarkably, there is a sharp decline in growth rate with increasing age, particularly for 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 start small and expand their assets as they mature, leading to a negative correlation between age and firm size (see Clementi and Hopenhayn (2006)). However, this relationship varies, as not all firms begin small and some remain small throughout their lifecycle.

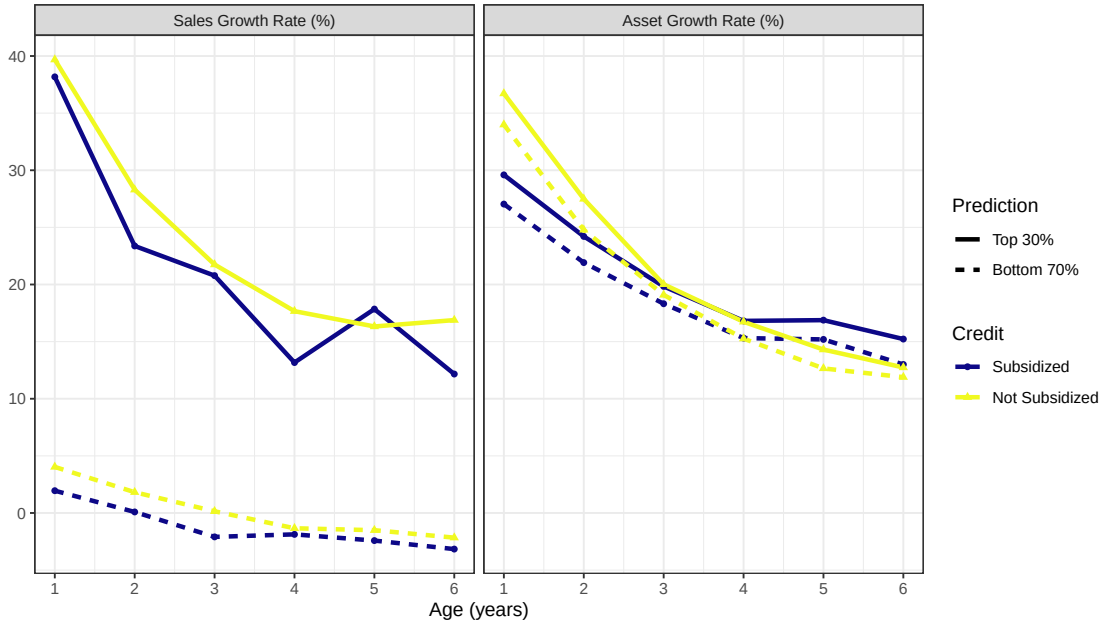

Figure 5: Firm performance comparison by ML prediction and age.
Notes: The figure presents the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target and age (defined at each year).

We subdivided firms into size and age and compared their 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. In the smallest size group, the negative correlation between growth rate and age is much stronger for subsidized firms compared to non-subsidized ones. 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. This implies that such firms may not face significant

credit constraints, even though they have received subsidized loans.

Efficiency in the subsidy loan program could be significantly enhanced by considering firm age more comprehensively, beyond just using it as an eligibility criterion. Program managers would be better off paying 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).

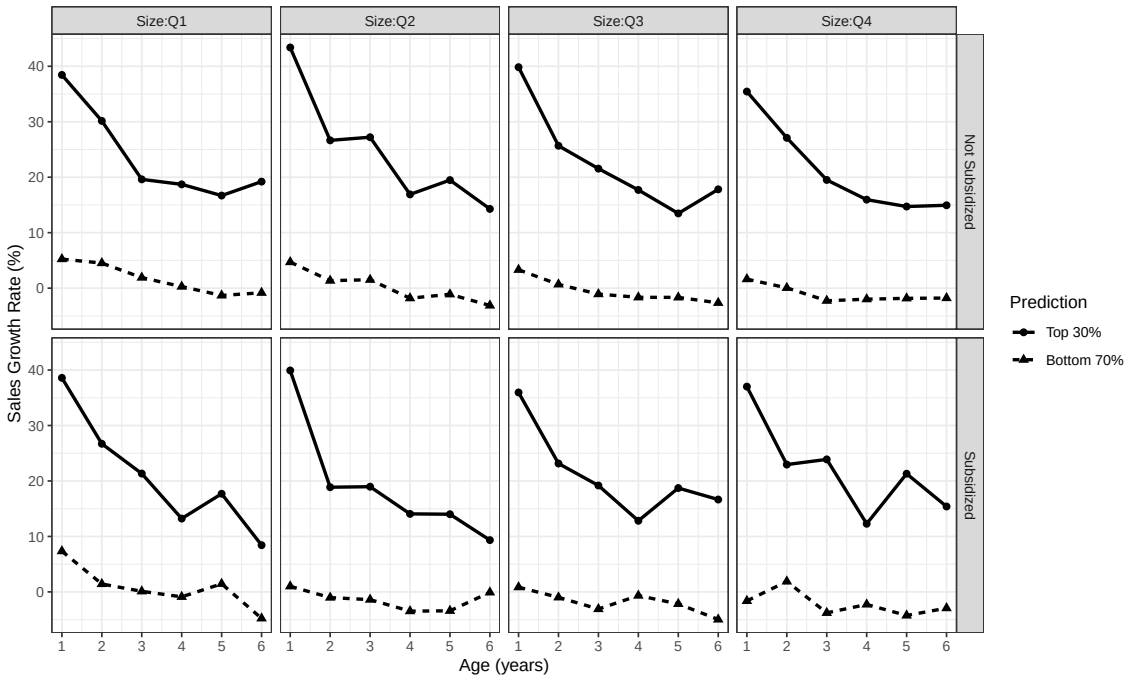

Figure 6: Firm performance comparison by ML prediction, size, and age.
Notes: The figure shows the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target, age, and asset size quantiles defined at each year and industry group.

We now explore the debt-to-capital ratio and cash flow ratio as indicators of credit constraints to identify potential targets for policy intervention. 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.

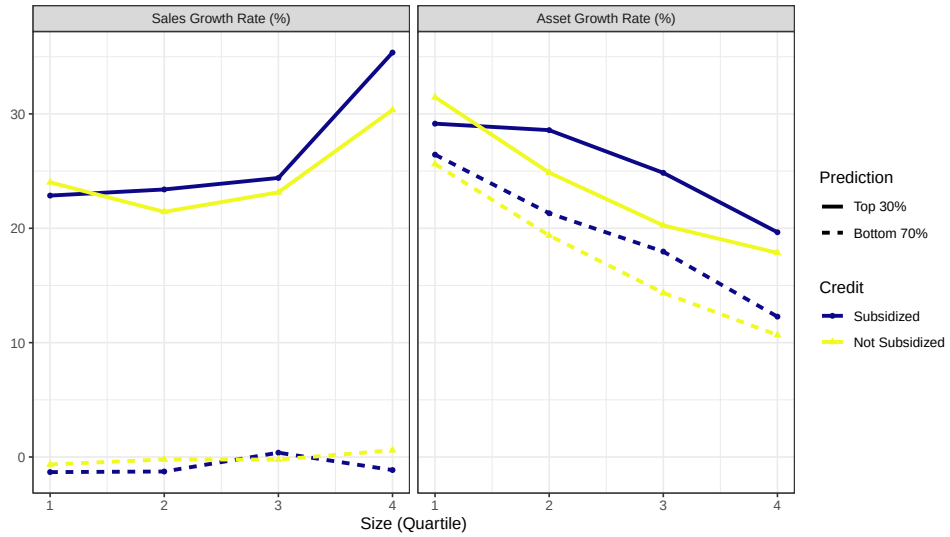

Figure 7: Firm performance comparison by ML prediction and debt-to-capital ratio
Notes: The figure displays the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target and quantiles of debt-to-capital ratio defined at each year and industry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

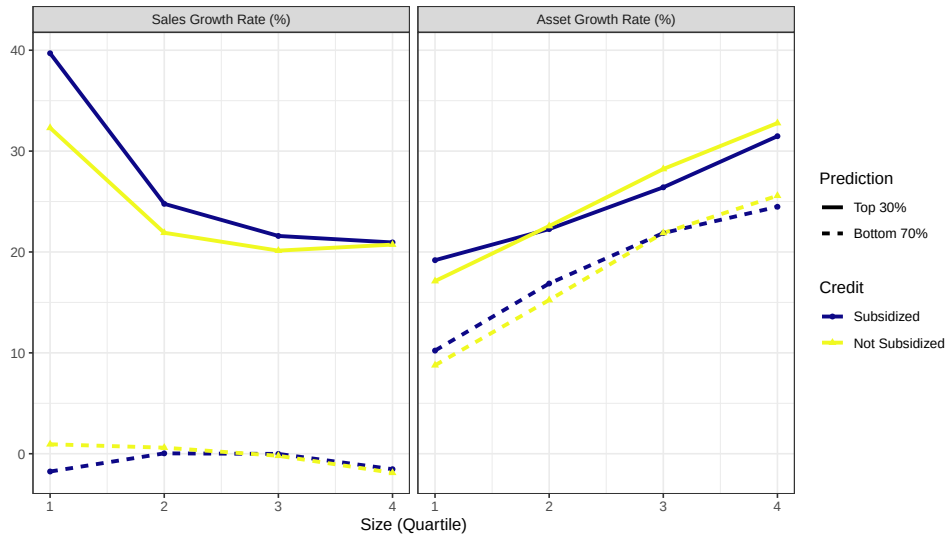

Figure 8: Firm performance comparison by ML prediction and cash flow ratio
Notes: The figure displays the average growth rates of sales and total assets for subsidized and non-subsidized firms, categorized by ML predicted target and quantiles of cash flow ratio defined at each year and industry group.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

by deliberately inflating expenses or writing off inventory 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. Utilizing an administrative database containing comprehensive records of 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 demonstrating the effectiveness 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% of the predicted sales growth group exhibited negligible average sales growth, and the lowest 30% of the predicted group experienced negative growth. There may be tied-up resources that could be more effectively reallocated to more productive and viable enterprises. 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 offers broad applicability across various public business support programs for identifying suitable targets. The future growth predictions can be utilized as either a supplementary or primary criterion during 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.[16] We found that the sales growth rates of firms with

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

low predicted growth were even lower for the subsidized group than not subsidized group. The credit program managers might have had objectives beyond just the growth prospects of firms, including preventing defaults and the associated job losses. However, given the primary objectives of government credit support policies, these programs should target firms that exhibit strong growth potential upon receiving credit. This approach allows policymakers to more effectively assess the programs' success in meeting their intended 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 struggling firms, policymakers can reduce inefficiencies and minimize the impact of 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 a firm's future performance relative to other firms within the same industry, it does not guide 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 economic impact of industries when making decisions about credit allocation. Consequently, this might necessitate the development of an additional model to address the credit allocation problem between industries.

Third, to develop an effective decision aid utilizing the ML prediction algorithm, a rigorous trial process and evaluation are essential. Relying solely on ML predictions for determining eligibility criteria may pose challenges due to potential significant variations in sales growth rates. Moreover, interpreting the decisions made by the model can be challenging, given its 'black box' nature. Therefore, further research could explore ways to enhance model interpretability, aiming to provide readily understandable outcomes.

One possible approach is to introduce a two-step verification process for firms whose sales

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

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. Through a rigorous trial and post-evaluation process, the ML algorithm can be refined to serve as a more effective decision aid.

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.

Baqaee, David Rezza and Emmanuel Farhi, "Productivity and misallocation in general equilibrium," *The Quarterly Journal of Economics*, 2020, *135* (1), 105–163.

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.

Caballero, Ricardo J, Takeo Hoshi, and Anil K Kashyap, "Zombie lending and depressed restructuring in Japan," *American economic review*, 2008, *98* (5), 1943–1977.

Chaney, Paul K and Anjan V Thakor, "Incentive effects of benevolent intervention: The case of government loan guarantees," *Journal of Public Economics*, 1985, *26* (2), 169–189.

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.

Coad, Alex and Stjepan Srhoj, "Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms," *Small Business Economics*, 2020, *55* (3), 541–565.

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.

Gale, William G, "Federal lending and the market for credit," *Journal of Public Economics*, 1990, *42* (2), 177–193.

_ , "Economic effects of federal credit programs," *The American Economic Review*, 1991, pp. 133–152.

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.

Horvath, Akos and Peter Lang, "Do loan subsidies boost the real activity of small firms?," *Journal of Banking & Finance*, 2021, *122*, 105988.

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.

Innes, Robert, "Investment and government intervention in credit markets when there is asymmetric information," *Journal of Public Economics*, 1991, *46* (3), 347–381.

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

Table A.I: Predictor rankings by their variable importance

Variable Name	Average Rank	Number of times in Top 5 Rankings	Percentage of times in Top 5 Rankings
Sales	1.67	70	97.22%
Lag of sales	3.15	62	86.11%
Change in sales	4.36	57	79.17%
Markup	5.72	34	47.22%
Change in liability	7.07	30	41.67%
Change in total assets	7.97	21	29.17%
Operating income	8.14	24	33.33%
Lag of operating income	8.43	14	19.44%
Cash flow ratio	8.82	5	6.94%
Change in operating income	9.42	4	5.56%
Number of employees	9.83	29	40.28%
Debt-to-capital ratio	11.04	2	2.78%
Tangible asset	11.29	4	5.56%
Lag of debt-to-capital ratio	11.76	3	4.17%
Lag of total assets	14.44	0	0%
Lag of liability	15.65	0	0%
Total assets	16.86	0	0%
Liability	17.58	0	0%
Paid-in capital	18.36	0	0%
Age	18.49	1	1.39%
Company size	21.08	0	0%
Govt subsidy dummy	22.26	0	0%
Venture firm status	23.12	0	0%
Type of company	23.46	0	0%

Note: The table reports the average rankings of variables for each industry-year block by their variable importance. It also presents the frequency and the percentage of times wherein the variable is ranked among the top five predictors in terms of variable importance. The term 'lag of a variable' refers to the 1-year lagged value of the variable. And 'the change of a variable' is defined as the logarithmic difference between the current value of the variable and its value at the 1-year lag.

Table A.II: 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.