Sfide metodologiche per la previsione del successo di imprese innovative

Gioia Di Credico

Abstract

Il lavoro si pone l'obiettivo di studiare ed identificare quali aziende giovani realizzeranno una forte crescita nei loro primi 5 anni di vita a partire dalle informazioni disponibili dopo il loro primo anno di vita. Un ampio dataset, ottenuto dalla banca dati Orbis, è stato utilizzato per le analisi. Le informazioni disponibili riguardano dati relativi al profilo economico, innovativo e strutturale delle aziende. La definizione scelta di forte crescita identifica circa il 3% delle giovani aziende presenti nel dataset come tali. Questo comporta la necessità di bilanciare i dati per poter ottenere risultati predittivi migliori. I modelli statistici evidenziano l'importanza della posizione geografica e permettono di evidenziare quali variabili hanno maggior impatto sulla forte crescita delle giovani aziende.

Keywords

Aziende ad alta crescita, start-up, classi sbilanciate, ROSE, GLMM

Profilo biografico

Gioia di Credico è ricercatrice a tempo determinato di tipo A per il settore SECS-S/01 presso il Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti" ed è stata assegnista di ricerca presso lo stesso dipartimento. Ha

ottenuto il titolo di dottore di ricerca presso l'Università degli Studi di Padova nel 2018. I suoi principali interessi di ricerca si concentrano sullo sviluppo ed sull'applicazione di metodologie semiparametriche, modelli gerarchici Bayesiani ed analisi di dati sbilanciati. È docente a contratto ed esercitatore presso l'Università degli Studi di Trieste in diversi insegnamenti di statistica per corsi di laurea triennale e magistrale.

1. Introduzione

Una start-up per definizione è un'azienda giovane a cui è associato un elevato tasso di rischio ed un basso tasso di sopravvivenza. Tra queste, alcune realizzeranno una forte crescita. L'obiettivo di questo lavoro è di prevedere quali start-up cresceranno fortemente sulla base delle loro caratteristiche rilevate nel primo anno di vita. Questo ambizioso proposito punta ad avere vantaggi economici. Infatti individuare le aziende a forte crescita sin dal loro primo anno di vita, porterebbe ad una riduzione degli elevati rischi di investimento e produrrebbe elevati guadagni in caso di previsione corretta. A questo si lega una maggiore consapevolezza del profilo delle aziende a forte crescita, le quali risultano di grande interesse non solo perché sono tra le principali fonti di creazione di lavoro, ma anche perché sono caratterizzate da un'elevata produttività che può arrivare a migliorare le performance di un intero Paese. Noti esempi di aziende a forte crescita sono JustEat, Aruba o Dropbox. Da un punto di vista metodologico le principali problematiche da affrontare sono state due: il forte sbilanciamento presente nella classe di interesse e l'elevato numero di variabili fortemente correlate tra loro. Mentre da un punto di vista pratico, il lungo processo di pulizia ed esplorazione dei dati e l'elevata complessità computazionale dei modelli sono stati i due fattori che hanno richiesto una maggiore quantità di tempo. Il lavoro è così organizzato: nella sezione "Materiali" vengono presentati i dati, come sono stati raccolti, puliti ed analizzati da un punto di vista descrittivo; la sezione "Metodi" descrive i modelli statistici applicati e le soluzioni proposte per le problematiche riscontrate; la sezione "Risultati" presenta i risultati dei modelli stimati ed in ultimo troviamo le conclusioni e spunti per possibili sviluppi futuri.

2. Materiali

I dati. I dati utilizzati sono stati scaricati dal database Orbis della società Bureau van Dijk [1] il quale rappresenta la più grande e completa fonte di informazione a livello mondiale sulle aziende, quotate e non. Le fonti che concorrono al popolamento della base dati sono 170 provider internazionali

e diverse fonti interne. Gli ambiti descritti per ciascuna azienda riguardano principalmente aspetti finanziari, espressi attraverso i dati di bilancio di esercizio, corredati da informazioni su altri ambiti, quali ad esempio la posizione, il numero di impiegati o il numero di brevetti. In questo modo si riesce ad avere una visione completa delle società presenti nel database. La società Bureau van Dijk si occupa non solo della raccolta ma anche della standardizzazione dei formati, passaggio imprescindibile per garantire la confrontabilità dei dati raccolti da varie fonti, e della stima di modelli finanziari che aiutano nella valutazione della solidità di un'azienda.

I dati scaricati rispondono ad un profilo aziendale preciso ottenuto combinando diversi criteri di ricerca disponibili su Orbis. Le aziende selezionate sono tutte nate nel 2010, attive al 2019 e con forma legale standardizzata attiva. Il fatturato (o turnover), definito come le entrate al netto delle spese operative, è una variabile di bilancio di particolare rilevanza per la nostra analisi, per questo motivo non deve essere mancante almeno nei primi due anni considerati (2010 e 2011). Altre variabili che devono avere valore non mancante relativamente al 2010 sono: gli asset totali, i fondi degli azionisti ed il settore NACE Rev. 2. Le informazioni disponibili riguardano 151.396 aziende e sono registrate in 249 variabili che descrivono: variabili di bilancio dal 2010 al 2014, variabili di innovazione ed altre informazioni descritte più nel dettaglio in seguito.

Preparazione dei dati. Si è reso necessario un attento processo esplorativo dei dati volto alla loro pulizia e selezione. In particolare sono state eliminate le aziende non attive nel periodo di osservazione 2010-2014 (circa 4.000), quelle che presentavano la parola 'holding' (o derivate) nel nome (circa 1.000), le società con fatturato superiore al miliardo (25) o con fatturato mancante o negativo nel periodo di osservazione 2010-2014 (circa 28.000). Anche le aziende classificate come grandi e molto grandi, cioè con ricavi di esercizio maggiori di 10 milioni di euro, con più di 150 dipendenti o patrimonio totale maggiore di 20 milioni di euro (circa 3.000 aziende) non sono state considerate. In ultimo, abbiamo esplorato la distribuzione delle aziende nei diversi Stati, rilevando che sia il numero di aziende sia il livello di dettaglio delle informazioni presenti sono influenzati dagli obblighi di legge di pubblicazione del bilancio di esercizio vigenti in ogni Paese. Stati con molte aziende presenti nel dataset ed informazioni dettagliate sono ad esempio l'Italia e la Francia, mentre per la Svizzera il numero di società presenti è molto limitato. Con l'obiettivo di esaminare anche l'effetto della posizione geografica sulla forte crescita delle giovani

aziende, si è deciso di non considerare quelle in Paesi con numerosità minore di 100 (circa 150 aziende in totale).

Come già detto precedentemente, il lavoro si pone l'obiettivo di prevedere quali aziende saranno caratterizzate da una forte crescita utilizzando dati relativi al loro primo anno di vita. Per questo, gli step successivi del progetto si concentrano esclusivamente sulle informazioni presenti nel dataset raccolte nel 2010. Inoltre, la forte dipendenza temporale delle variabili di bilancio appartenenti ad anni consecutivi, se non modellata correttamente, potrebbe portare ad avere risultati poco precisi in alcuni tipi di modelli statistici. Riguardo al settore, nel dataset sono presenti due classificazioni: la classificazione NACE Rev. 2, che rappresenta un sistema di classificazione statistica internazionale della Comunità Europea, e la classificazione della società Bureau van Dijk. La classificazione NACE Rev. 2 presenta diversi livelli di dettaglio (21 settori e 732 sotto-settori) non disponibili nella classificazione di Bureau van Dijk (19 settori). Le due classificazioni primarie risultano praticamente equivalenti ma decidiamo di mantenere la classificazione NACE Rev. 2 in vista di un possibile ampliamento dell'analisi considerando categorie più specifiche.

L'esplorazione dei dati ha evidenziato una quantità importante di dati mancanti su diverse variabili. L'esclusione diretta di tutte aziende con voci incomplete avrebbe portato ad una perdita molto forte di informazioni riducendo il dataset a circa 30.000 voci e limitando la possibilità di stima di modelli complessi. L'imputazione dei dati mancanti al contrario permette di recuperare delle informazioni, ma questa operazione, se non condotta con attenzione, può condizionare i risultati delle analisi. Per ridurre questo rischio si è scelto di imputare i soli valori che possano essere ricostruiti uguali a 0. A tal fine abbiamo sfruttato le semplici operazioni algebriche che legano tra loro le variabili di bilancio. Invertendo queste semplici relazioni possiamo infatti verificare se alcuni valori mancanti possono essere imputati come 0. Se la differenza tra la variabile ricostruita e quella presente nel dataset è minore di 10.000 euro in valore assoluto, imputiamo come 0 i valori mancanti delle variabili utilizzate nel calcolo, altrimenti eliminiamo l'azienda con il valore mancante. Ad esempio gli asset fissi sono calcolati come la somma di asset fissi intangibili, asset fissi tangibili ed ulteriori asset fissi. Ognuna delle ultime tre variabili presenta circa 8.000 valori mancanti che, se imputati come 0, ci permettono di ricostruire la variabile asset fissi correttamente ad eccezione di 20 aziende eliminate dal dataset. Tuttavia, quando il numero di valori mancanti è elevato o quando alcune variabili necessarie nel calcolo delle relazioni di bilancio

non sono disponibili nel dataset, le variabili sono state escluse. È il caso ad esempio della variabile debiti a breve termine (current liabilities) che è stata considerata da sola nell'analisi, data l'assenza della variabile prestiti (loans) e l'elevato numero di valori mancanti sulle variabili creditori (creditors) e altri debiti a breve termine (other current liabilities). Tramite questo passaggio abbiamo recuperato una quantità importante di dati e contemporaneamente abbiamo svolto una prima selezione delle variabili utilizzabili nei modelli.

L'ultimo aspetto considerato nel processo di pulizia dei dati riguarda la verifica della presenza di valori anomali, cioè valori molto grandi o molto piccoli rispetto alla tendenza centrale delle variabili o fuori dallo spettro dei valori possibili. Ad esempio, la variabile numero di società nel gruppo aziendale presenta il valore mediano pari a 2, il terzo quartile pari a 4 ed un valore massimo pari a 1.710. Per valutare se l'utilizzo di dati più puliti comporti un miglioramento in termini di stima del modello, abbiamo eseguito le analisi statistiche in parallelo su tre dataset: il primo completo, composto da 110.414 aziende; il secondo che esclude lo 0.1% dei valori estremi su ciascuna delle variabili numeriche, composto da 108.346 aziende; ed il terzo che esclude l'1% dei valori estremi, composto da 96.834 aziende.

La variabile risposta. Le variabili di cui ci siamo occupati fino ad ora descrivono aspetti che possono aiutarci nella previsione dello stato di forte crescita di un'azienda. Di seguito ci riferiremo a queste come variabili esplicative. Mentre, la variabile, cosiddetta, di risposta, cioè quella che descrive lo stato di forte crescita di un'azienda nei suoi primi cinque anni di vita, non è presente nel dataset ottenuto da Orbis ed è stata costruita da noi. Esistono varie definizioni di azienda a forte crescita che portano a classificazioni diverse. Ne abbiamo valutate tre, tutte basate sui valori della variabile fatturato nel periodo 2010-2014. La prima definizione di forte crescita è quella proposta dall'Eurostat che classifica un'azienda come a forte crescita se registra una crescita media annua maggiore o uguale al 20% su un periodo di almeno tre anni. Questo criterio ci porta ad individuare circa il 40% delle aziende come a forte crescita. La seconda definizione valutata è quella del Compound Annual Growth Rate (CAGR) secondo cui un'azienda a forte crescita è caratterizzata da uno sviluppo medio nei primi 5 anni maggiore o uguale al 20% e fatturato maggiore di 1 milione. Essendo un criterio più rigido del precedente perché basato su un periodo più ampio e con vincolo sul fatturato, solo il 10% delle aziende vengono individuate come aziende a

forte crescita secondo questa definizione. In ultimo, abbiamo considerato la definizione proposta da Birch et al. [2] secondo cui è necessaria una crescita maggiore del 20% annuo per un periodo di almeno 5 anni per poter classificare un'azienda come "gazelle". Questa definizione, la più rigida delle tre poiché pone un vincolo di crescita minima nei primi cinque anni, porta ad identificare solo il 3% delle aziende come a forte crescita. Decidiamo di utilizzare questa definizione nelle analisi successive perché in linea con le informazioni reperibili in letteratura. Inoltre descrive il profilo di crescita aziendale a cui siamo maggiormente interessati.

Analisi descrittiva. Esaminiamo ora più nel dettaglio le variabili selezionate e la loro relazione con la variabile risposta. Nel dataset risultano circa 40 variabili esplicative, di cui 23 legate al bilancio di esercizio del 2010. Nella figura 1 sono rappresentate le correlazioni tra le variabili di bilancio relative al 2010 ed emerge come le variabili maggiormente correlate siano quelle legate per definizione da relazioni lineari. Questa informazione va tenuta a mente nella valutazione dei modelli e nella futura selezione delle variabili. Il secondo insieme, composto da 6 variabili, descrive il carattere innovativo delle aziende; troviamo variabili come il numero di licenze, brevetti e marchi registrati. In letteratura, questo sembra essere un aspetto rilevante nella crescita delle aziende giovani. Notiamo come solo circa lo 0.6% delle aziende risulta avere almeno una delle suddette caratteristiche innovative e di queste circa l'8% risulta essere un'azienda a forte crescita. La distribuzione delle aziende per settore non è uniforme. Un quarto delle aziende del dataset è classificato come "G - Commercio all'ingrosso e al dettaglio; riparazione di autoveicoli e motocicli", ed in questo settore troviamo l'1% di tutte le aziende classificate come a forte crescita. Altri settori molto rappresentati sono il settore "F - Costruzioni" (12.5%) e quello "M - Attività professionali, scientifiche e tecniche" (12.5%). Il 7% delle aziende nella categoria "E - Fornitura di acqua; reti fognarie, attività di trattamento dei rifiuti e risanamento" ed il 6% nella categoria "H – Trasporti e magazzinaggio" e "P - Istruzione" risultano aziende a forte crescita. La posizione geografica è espressa in Nazione, città e codice postale. Quattro Stati su un totale di 24 contribuiscono al 60% dei dati disponibili. Infatti, circa un quarto delle aziende presenti nel dataset si trova in Italia, seguono la Romania (circa il 14%), la Francia (circa l'11%) e la Russia (circa il 10%). La figura 2 mostra la distribuzione delle aziende nei vari Stati. È interessante notare poi come in alcuni Paesi con un basso numero di aziende registrate (meno di 1.000), la percentuale di quelle ad alta

crescita è più elevata rispetto al quella degli stati più rappresentati: è il caso di Gran Bretagna, Polonia, Bosnia Erzegovina, Lussemburgo e Turchia. Un'altra variabile rilevante nella valutazione della crescita di un'azienda è il numero di impiegati. Nel nostro dataset, il numero medio di impiegati è pari a 3.28 mentre i primi due quartili risultano pari a zero ed il terzo pari a 2. Ulteriori informazioni a disposizione riguardano la forma legale dell'azienda, la categoria della società, il numero di azionisti, il numero di aziende sussidiarie o nello stesso gruppo, l'indipendenza rispetto agli azionisti ed il livello di consolidamento finanziario.

3. Metodi

Modelli statistici. I modelli statistici sono strutture matematiche che descrivono la relazione assunta tra la variabile risposta e le esplicative sui dati osservati, ipotizzando un meccanismo generatore. Esiste un'ampia gamma di modelli, caratterizzati da diversi livelli di complessità. I più semplici in genere descrivono dinamiche ben definite e facilmente interpretabili, ma possono non essere la miglior opzione in termini di accuratezza predittiva. I più complessi sono spesso caratterizzati da ipotesi di relazioni di difficile interpretazione ma grande accuratezza accompagnata da un'elevata complessità computazionale. Lo scopo dell'analisi dovrebbe aiutare nella scelta del modello che si vuole stimare. Ad esempio, se si è più interessati a capire quali fattori influiscono maggiormente sulla forte crescita di un'azienda ci si può orientare verso un modello di regressione, mentre se l'obiettivo è quello di riuscire ad ottenere previsioni accurate, rilassando ipotesi stringenti sulle relazioni tra le variabili esplicative e quella di risposta, si possono scegliere modelli complessi quali metodi di ensemble o reti neurali. Ovviamente non è detto che modelli semplici non possano arrivare a capacità predittive di modelli più complessi, molto dipende dalla reale natura della relazione in esame e da quanto il modello è capace di approssimarla e descriverla.

Nel nostro lavoro abbiamo considerato diversi modelli con il duplice scopo di spiegare la relazione tra la forte crescita delle aziende giovani e le variabili a disposizione nel dataset e prevedere con accuratezza la forte crescita. A tale scopo abbiamo stimato i seguenti modelli statistici: un modello logistico, un albero di classificazione, un modello logistico gerarchico Bayesiano che include la posizione geografica di ogni azienda e ci permette di valutarne l'impatto, ed infine una rete neurale. Il modello logistico mette in relazione una trasformazione della probabilità di essere

a forte crescita, con le variabili esplicative [3]. In particolare il logaritmo dell'odds di essere un'azienda a forte crescita viene modellato linearmente attraverso le variabili esplicative. Il modello logistico si colloca nella famiglia dei modelli lineari generalizzati che include diverse possibili estensioni. Una di queste è quella di poter considerare la struttura a gruppi dei dati (paesi) attraverso l'inclusione nel modello di effetti casuali [3]. L'estensione Bayesiana [4] poi considera informazioni sul fenomeno disponibili a priori che vengono aggiornate attraverso l'evidenza empirica. L'approccio permette di rappresentare anche situazioni di non conoscenza del fenomeno a priori, nel nostro caso sono state incluse informazioni a priori vaghe e solo relative ai valori matematici plausibili dei parametri. Sempre attraverso la definizione delle informazioni a priori, è stato stimato un modello logistico gerarchico Bayesiano con selezione delle variabili esplicative che hanno un impatto significativo sulla risposta. Un modello diverso è rappresentato dall'albero di classificazione [5]. Attraverso un processo ricorsivo binario, ad ogni iterazione viene individuata una bipartizione basata sul valore di una variabile esplicativa che ottimizza un determinato criterio. Il dataset viene così diviso in due parti ed il processo prosegue sui due sotto insiemi ottenuti. Diversi parametri regolano la crescita dell'albero, cioè il numero di divisioni. Tuttavia è buona pratica lasciar crescere l'albero e successivamente potarlo al livello che minimizza l'errore di validazione incrociata. Infatti un albero troppo grande può portare ad un modello sovrastimato, cioè un modello troppo aderente ai dati osservati e poco generalizzabile. Un'estensione di questo metodo è un metodo di ensemble chiamato random forest che combina diversi alberi di classificazione. Attraverso questo metodo è possibile ottenere una classificazione delle variabili più importanti [6]. Tra quelle testate, le reti neurali rappresentano la metodologia più flessibile ma anche meno interpretabile. La stima del modello infatti è costituta da diversi livelli nascosti in cui sono presenti dei nodi costituiti da combinazioni delle variabili. Sfortunatamente la complessità computazionale richiesta da questi metodi unita all'ampiezza del nostro dataset, ci ha permesso di poter stimare solamente reti neurali molto semplici non adatte allo studio del fenomeno in esame.

Le analisi esplorative, i grafici ed i modelli sono stati eseguiti con il software R [7].

ROSE. Solo il 3% delle aziende che sopravvivono ai primi 5 anni di vita diventano aziende a forte crescita. Questo sbilanciamento rende la stima

del modello con tecniche statistiche standard molto complessa poiché le informazioni caratterizzanti la classe minoritaria risultano di difficile apprendimento da parte del modello. Può capitare infatti che il processo di stima consideri come rumore quello che è per noi l'obiettivo dell'analisi. Tra le varie tecniche di bilanciamento dei dati proposte in letteratura troviamo il Random Over Sampling Examples (ROSE) [8]. ROSE è una strategia ibrida capace di superare i limiti del sovra-campionamento, che prevede la ripetizione degli stessi dati osservati fino al raggiungimento della numerosità desiderata, e quelli del sotto-campionamento, che porta ad una forte perdita di informazioni. Infatti, ROSE permette di ottenere dataset bilanciati rispetto alla variabile risposta, quindi il dataset bilanciato avrà la stessa dimensione del dataset in entrata ma, rispetto alla variabile risposta, sarà composto per il 50% da aziende a forte crescita. Inoltre, le nuove osservazioni sulle variabili esplicative sono generate a partire dalla stima di una densità kernel, con il vantaggio di simulare valori non esattamente uguali a quelli osservati ma coerenti con la distribuzione della variabile.

Misure di accuratezza. Per poter valutare le capacità predittive dei modelli, è utile impiegare dati esterni rispetto a quelli impiegati nello step di stima. A tal proposito abbiamo diviso in modo casuale il dataset in due parti, la prima è detta training set, contiene l'80% dei dati e verrà usata nella stima del modello; mentre la seconda parte, di dimensione pari al 20% dei dati, è stata utilizzata come test set. I risultati ottenuti, ovvero i valori predetti relativamente alle unità del test set, vengono poi confrontati con i valori osservati presenti nel test set ed inseriti nella matrice di confusione in tabella 1 [9]. Definiamo così quattro possibili combinazioni: i veri positivi (TP), cioè i valori osservati come a forte crescita e predetti correttamente, i veri negativi (TN), i valori osservati come non a forte crescita e predetti correttamente, i falsi negativi (FN), valori predetti come non a forte crescita che però lo sono, ed i falsi positivi (FP), valori predetti come aziende ad alta crescita che non lo sono. Da queste quattro quantità è possibile costruire diverse metriche di validazione della capacità predittiva del modello. La metrica di validazione più utilizzata è l'accuratezza definita come il rapporto tra i totali previsti correttamente (TP+TN) ed il totale dei valori predetti (TP+TN+FN+FP). L'accuratezza misura la proporzione di valori predetti correttamente, non differenziando i veri positivi ed i veri negativi. Nel nostro caso tuttavia, tenendo conto del forte sbilanciamento delle due classi, queste due quantità andrebbero valutate in modo più puntuale. Infatti, nel caso peggiore rispetto alla categoria a forte crescita, ovvero

se tutte le aziende di questa classe venissero predette in modo errato, si otterrebbe comunque un'accuratezza molto alta pari alla proporzione della classe più rappresentata, nel caso in analisi uguale al 97%. Altre misure che si possono calcolare sono la precisione, la sensibilità e la specificità. La prima è definita come il rapporto tra i veri positivi (TP) ed il totale dei positivi predetti (TP + FP) e rappresenta la proporzione di aziende a forte crescita predette correttamente sul totale delle aziende predette in questa categoria. La sensibilità misura la proporzione di aziende ad alta crescita correttamente predette (TP) rispetto al totale delle aziende ad alta crescita osservate (TP + FN). È una misura particolarmente rilevante nella nostra analisi poiché quando è alta identifica un modello in grado di individuare correttamente le aziende a forte crescita con un rischio basso di classificarle in modo errato. In ultimo definiamo la specificità come il rapporto tra i veri negativi (TN) ed il totale dei valori osservati come non a forte crescita (FP + TN). Questa misura descrive la proporzione di aziende non a forte crescita che vengono correttamente classificate come tali.

4. Risultati

In questa sezione presentiamo i risultati per i modelli stimati. Come anticipato nella sezione relativa ai dati, le analisi sono state condotte su tre dataset. Tuttavia i risultati riportati si concentrano su quelli ottenuti con il dataset che esclude lo 0.1% dei valori estremi sulle variabili numeriche. Infatti, comparando i modelli in termini di velocità di stima e performance predittive abbiamo notato un peggioramento utilizzando il dataset completo rispetto ai dataset con il taglio sui valori estremi. Mentre i risultati numerici sono equivalenti tra i tre dataset. Volendo sfruttare il maggior numero di informazioni disponibili, abbiamo reputato conveniente presentare le stime per il dataset di dimensioni intermedie.

I modelli sono stimati sia sul training set bilanciato con ROSE (50% aziende a forte crescita; 50% aziende non a forte crescita) che sul training set sbilanciato (3% aziende a forte crescita; 97% aziende non a forte crescita), mentre il test set è sempre sbilanciato.

I risultati delle misure di accuratezza sono raccolti nella tabella 2. Riguardo al modello logistico, le variabili significative sono: il Paese, il settore NACE Rev. 2 (21 categorie), le variabili di bilancio assets, stock, liabilities e turnover, il numero di impiegati, il numero di azionisti e le informazioni sulle aziende sussidiarie. Sia il modello stimato sul training set sbilanciato che quello sul training set bilanciato riescono ad identificare molte aziende ad alta crescita

con una sensibilità circa del 60%, ma il numero di falsi positivi è purtroppo molto alto, infatti la precisione è appena del 7%. L'albero di classificazione risulta in linea con il modello logistico riguardo le variabili più influenti, tuttavia si evidenzia una difficoltà del modello nel prevedere le aziende a forte crescita con una sensibilità pari a 0 sia nel caso del training set sbilanciato che in quello del training set bilanciato. Nel caso del modello logistico gerarchico Bayesiano si nota un'evidente differenza tra il modello stimato sui dati non bilanciati, che tende a prevedere molti falsi positivi con una precisione del 3%, ed il modello su dati bilanciati che risulta più equilibrato nelle previsioni, seppur con un elevato numero di falsi negativi evidenziato da una sensibilità del 21%. Questo modello conferma l'effetto significativo della posizione geografica già osservato nel modello logistico (primo punto nella figura 3), inoltre gli effetti random mostrati in figura 3 mostrano un'evidente variabilità tra Paesi. Nel grafico, la linea orizzontale sopra ogni punto rappresenta la variabilità della stima: i paesi con maggiore variabilità risultano essere quelli rappresentati da meno aziende. In ultimo il modello logistico gerarchico Bayesiano con selezione delle variabili stimato su dati bilanciati risulta analogo al precedente ma permette di effettuare una selezione delle variabili e valutare il loro impatto sulla forte crescita delle aziende. Più in dettaglio, le variabili che risultano poco significative e di cui si può valutare l'esclusione sono: la forma legale, l'indicatore di indipendenza rispetto agli azionisti di Bureau van Dijk ed il livello di consolidamento finanziario. Sempre da questo modello ricaviamo le variabili con impatto positivo sulla forte crescita che risultano essere per il settore NACE Rev.2 le categorie "Q - Sanità ed assistenza sociale", "P - Istruzione", "J - Servizi di informazione e comunicazione", "H - Trasporto e magazzinaggio", "G - Commercio all'ingrosso ed al dettaglio; riparazione di autoveicoli e motocicli". Anche il numero di aziende sussidiarie e filiali ed il numero di brevetti sembrano avere in media un effetto positivo sulla forte crescita. Al contrario le variabili con impatto negativo sono: i marchi registrati, il patrimonio complessivo ed i settori "O – Amministrazione pubblica", "F - Costruzioni" e "D – Fornitura di energia elettrica, gas, vapore ed aria condizionata".

5. Conclusioni

Nonostante la grande numerosità campionaria e l'elevata quantità di variabili esplicative, lo sbilanciamento delle classi rende difficile riuscire ad estrarre le informazioni riguardanti la classe di maggior interesse che è poco rappresentata. La tecnica di bilanciamento applicata migliora in alcuni casi le

performance predittive del modello rispetto alla classe minoritaria. Il taglio dei valori estremi permette sempre di ottenere, a parità di risultati, stime in tempi più brevi e in alcuni casi anche modelli migliori in termini predittivi. Il modello logistico gerarchico Bayesiano con selezione delle variabili ci ha permesso di individuare in un solo passaggio le variabili non significative e valutare l'impatto di quelle significative sulla forte crescita. In ultimo, è stata valutata la rilevanza della localizzazione geografica che, dato l'impatto riscontrato, suggerisce possibili estensioni in questa direzione. Ad esempio, una classificazione del territorio più sottile, come il sistema NUTS, potrebbe portare all'identificazione di distretti innovativi ad alto fattore di crescita. Un'altra estensione possibile vede l'inclusione di categorie più specifiche rispetto ai settori NACE Rev. 2, informazione già disponibile nel dataset.

I modelli richiedono comunque risorse computazionali elevate e questo limite non ci ha permesso di testare a fondo quello che sembra essere il miglior strumento in termini predittivi, le reti neurali. Un altro interessante sviluppo dell'analisi riguarda la stima del pattern di crescita delle aziende, sempre a partire dai dati relativi al loro primo anno di vita. Questo permetterebbe di avere una visione più dettagliata del periodo di vita a più alto rischio delle giovani aziende.

Bibliografia

[1] Bureau van Dijk <https://www.bvdinfo.com/en-gb/>; Sito consultato il 29/01/2022.

[2] D. L. Birch, A. Haggerty, W. Parsons, *Who's Creating Jobs?*, Cognetics Incorporated, 1995.

[3] A. Gelman, J. Hill, *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press, 2006.

[4] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, *Bayesian data analysis*, CRC press, 2013.

[5] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and regression tre- es*, CRC press, 1984.

[6] J. Friedman, T. Hastie, R. Tibshirani, *The elements of statistical learning*, New York: Springer series in statistics, 2001.

[7] R Core Team, R: *A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, 2022.

[8] G. Menardi, N. Torelli, T*raining and assessing classification rules with imbalanced data*, "Data mining and knowledge discovery", pp. 28(1), 2014, 92-122.

[9] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, 2003.

		Valori predetti	
Valori osservati		A forte crescita	Non a forte crescita
	A forte crescita	TP	FN
	Non a forte crescita	FP	TN

Tabella 1: matrice di confusione: rappresenta le frequenze rispetto alle due categorie a forte crescita (HGF), non ad alta crescita (NON HGF) per i valori osservati (sulle righe) e predetti (sulle colonne). A partire da queste quattro quantità è possibile costruire diverse misure di validazione del modello.

		Dati sbilanciati	Dati bilanciati
Modello logistico			
	Accuratezza	0.74	0.71
	Precisione	0.07	0.07
	Sensibilità	0.59	0.63
	Specificità	0.74	0.71
Albero di classificazione			
	Accuratezza	0.97	0.97
	Precisione	0.08	0.00
	Sensibilità	0.00	0.00
	Specificità	1.00	1.00
Modello logistico Bayesiano gerarchico			
	Accuratezza	0.15	0.87
	Precisione	0.03	0.06
	Sensibilità	0.00	0.21
	Specificità	0.13	0.89
Modello logistico Bayesiano gerarchico con selezione delle variabili			
	Accuratezza	0.97	0.86
	Precisione	0.06	0.06
	Sensibilità	0.00	0.21
	Specificità	1.00	0.88

Tabella 2: risultati ottenuti delle metriche di validazione dei modelli stimati. I risultati sono presentati per i modelli stimati validati sul dataset che esclude lo 0.1% dei valori estremi su ciascuna variabile numerica. Il cut-off sulla probabilità di essere HGF è stato fissato pari a 0.04.

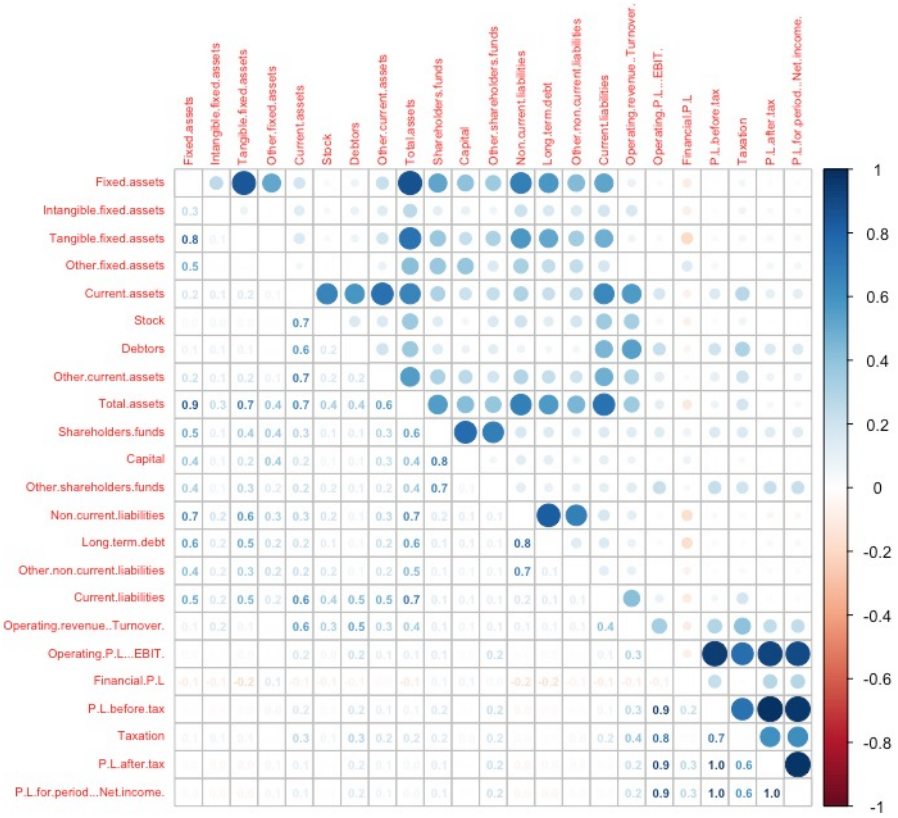

Figura 1: variabili di bilancio di esercizio relative al 2010, primo anno di vita delle aziende in esame. Si nota una forte correlazione positiva soprattutto tra le variabili legate, per costruzione, da relazioni lineari.

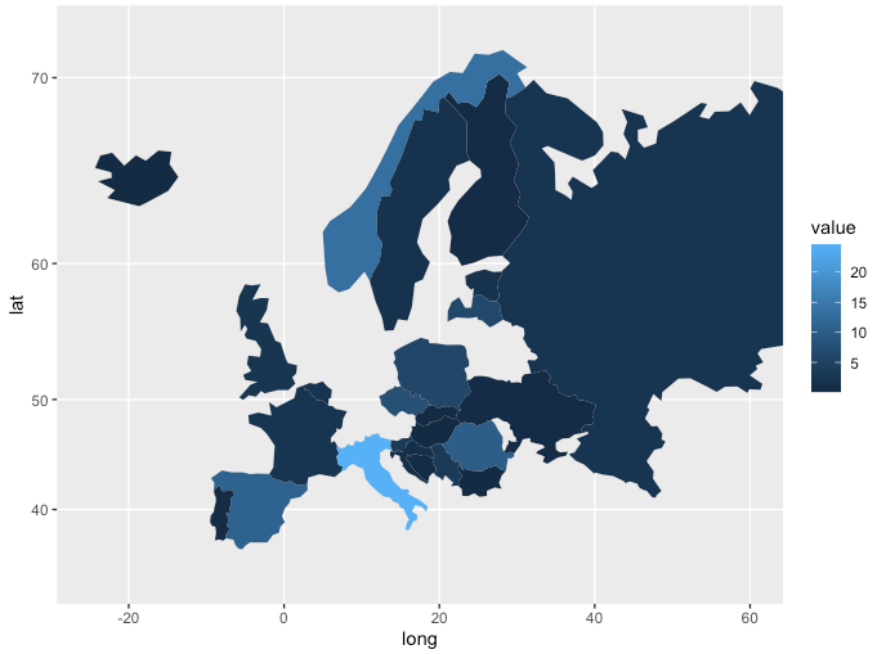

Figura 2: distribuzione delle aziende nate nel 2010 nei vari Paesi selezionati. I valori riportati nella legenda sono su scala percentuale. Il colore varia in base alla concentrazione: a colori più chiari corrispondono percentuali più alte e viceversa. Gli Stati esclusi non sono mostrati.

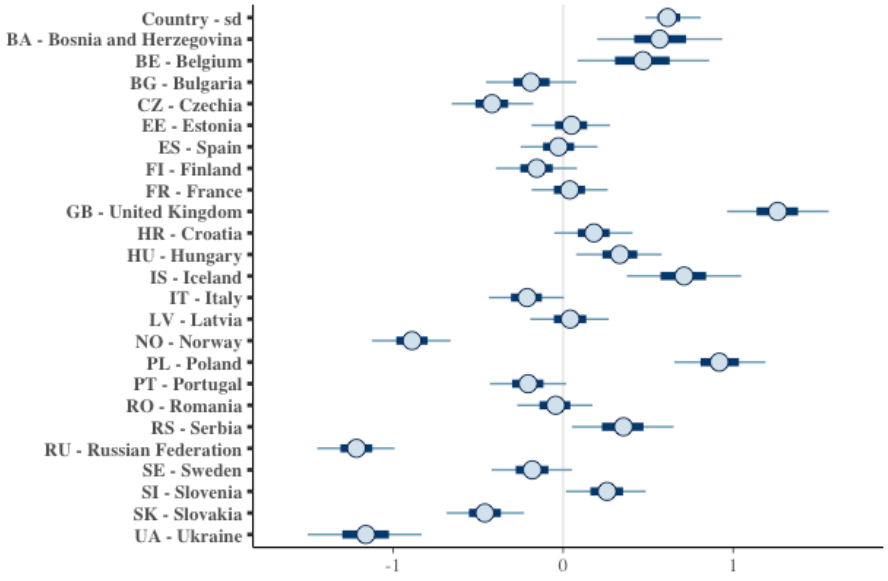

Figura 3: effetti random stimati dal modello logistico gerarchico Bayesiano stimato sul training test ottenuto dal dataset che esclude lo 0.1% dei valori estremi sulle variabili numeriche. La figura importa le mediane a posteriori degli intervalli di credibilità al 50% e al 90%.