Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in Scienze Statistiche

Text Mining nell'Analisi delle IPO: Uno Studio del Mercato Italiano dal 2010

Relatore: prof. **Andrea Sciandra**

Candidato: **Kevin Koci**

Matricola: **2062442**

Anno Accademico 2023/2024

Indice

Introduzione

L' Offerta Pubblica Iniziale, nota come IPO, rappresenta un bivio nel ciclo di vita di un'azienda. La scelta di quotarsi in borsa consente alle aziende di raccogliere capitale e aumentare la propria visibilità. Negli ultimi anni, il mercato italiano è stato caratterizzato da un aumento significativo delle aziende che hanno scelto di quotarsi in borsa.

L'obiettivo di questa tesi è analizzare l'underpricing delle IPO italiane. L'underpricing si riferisce alla differenza tra il prezzo di offerta delle azioni e il loro prezzo di chiusura nel primo giorno di negoziazione, una discrepanza che può fornire indicazioni preziose sull'efficienza del mercato e sulle aspettative degli investitori. Attraverso l'utilizzo di tecniche di text mining, questo studio si propone di esaminare l'influenza dei media tradizionali e dei social media sull'underpricing delle IPO italiane, con un focus sul periodo dal 2010 al 2022.

A sostegno di questo lavoro sono stati adottati diversi strumenti per la raccolta dei dati, tra cui CrowdTangle per i dati di Facebook, Google Trends per analizzare l'interesse di ricerca online, e web scraping per la raccolta di articoli da fonti finanziarie come SoldiOnline e MilanoFinanza. L'obiettivo è stato quello di costruire un dataset che permette di esplorare le relazioni tra le variabili economiche, le variabili riferite all'attenzione sui social media e media tradizionali e le performance delle IPO.

Nel capitolo 1 verranno trattate la struttura e il funzionamento dei mercati di

Borsa Italiana, l'ipotesi di efficienza del mercato, e l'influenza dei media e social media sull'underpricing delle IPO. Nel capitolo 2 verranno presentate le strategie utilizzate per la raccolta dei dati riferiti a Facebook, Google Trends, SoldiOnline e MilanoFinanza. Nel capitolo 3 verrà fornita una descrizione completa del dataset utilizzato nell'analisi, con un'analisi esplorativa per comprendere la distribuzione e le caratteristiche delle variabili. Il capitolo 4 presenta l'analisi e l'implementazione dei modelli di Machine Learning (Lasso, Random Forest e Support Vector Regression) utilizzati nello studio. Il capitolo 5 discute i valori SHAP (SHapley Additive exPlanation) per interpretare i modelli di Machine Learning utilizzati. Il capitolo 6 sintetizza i risultati ottenuti dai modelli e discute i limiti dello studio. Infine, il capitolo 7 conclude la tesi con una sintesi dei principali risultati.

Capitolo 1

Caratteristiche e influenze delle IPO italiane

1.1 Struttura e funzionamento dei Mercati di Borsa Italiana

La Borsa Italiana è gestita da Borsa Italiana S.p.A., parte del London Stock Exchange Group, e comprende sia mercati azionari regolamentati che non regolamentati. Come presentato da Teti e Montefusco(2021) ci sono due mercati regolamentati: il Mercato Telematico Azionario (MTA) e il Mercato Telematico degli Investment Vehicles (MIV).

Le azioni quotate sul Mercato Telematico Azionario (MTA) erano suddivise in tre categorie, determinate dalle dimensioni aziendali e dai requisiti specifici necessari per l'ammissione al segmento. Tali categorie erano denominate blue chip, star e standard. Successivamente, la distinzione tra le categorie blue chip e standard è stata eliminata, e i titoli sono stati classificati in base alla capitalizzazione e alla liquidità nei segmenti Large Cap, Mid Cap e Small Cap. Il segmento STAR, invece,

ha continuato a rappresentare le imprese di media capitalizzazione che aderiscono a rigorosi criteri di trasparenza, governance e liquidità.
Il MIV, invece, è il mercato regolamentato creato con l'obiettivo di fornire ai veicoli di investimento capitale, liquidità e visibilità per supportare la loro visione strategica. Per quanto riguarda i mercati non regolamentati, Borsa Italiana S.p.A. gestisce i Multilateral Trading Facilities (MTF), che includono il Global Equity Market (GEM), il Trading After Hours (TAH) e il Mercato Alternativo dei Capitali (AIM). L'AIM è il segmento di mercato riservato alle piccole e medie imprese (PMI) ad alto potenziale di crescita; è caratterizzato da un processo di ammissione semplice, pur garantendo un'alta visibilità internazionale. Questo segmento è stato creato con l'obiettivo di aiutare le PMI italiane nel loro processo di internazionalizzazione, fornendo loro le risorse finanziarie adeguate per competere contro gli avversari non italiani.
In base a quanto riportato dal sito di Borsa italiana, in seguito all'acquisizione di Borsa Italiana Spa da parte di Euronext, dal 25 ottobre 2021, i mercati equity di Borsa Italiana hanno cambiato denominazione:

- Il Mercato Telematico Azionario (acronimo MTA) è diventato Euronext Milan (nuovo acronimo EXM)

- Il Segmento STAR (acronimo STAR) è diventato Euronext STAR Milan (nuovo acronimo STAR)

- Il mercato AIM Italia (acronimo AIM) è diventato Euronext Growth Milan (nuovo acronimo EGM)

- Il mercato MIV (acronimo MIV) è diventato Euronext MIV Milan (nuovo acronimo MIV)

L'EXM e l'EGM sono i principali mercati per le IPO italiane. Le principali differenze tra questi due mercati riguardano il processo di ammissione, la governance aziendale, gli standard di trasparenza e divulgazione da mantenere come appartenenti a uno di questi segmenti, così come la liquidità degli asset. In particolare, le aziende che desiderano essere quotate sull'EXM devono avere una capitalizzazione minima di 40 milioni di euro e offrire una quota libera minima del 25%; devono essere state costituite da almeno tre anni e rispettare le regole di governance aziendale presentate nel Decreto Finanziario Italiano. Inoltre, queste aziende devono presentare alle autorità italiane competenti un documento dettagliato chiamato "prospetto", la cui veridicità è accertata attraverso un complesso processo di due diligence. Il prospetto include informazioni complete sull'emittente e sui titoli offerti, nonché un elenco dettagliato di tutti i rischi coinvolti sia nell'azienda che emette i titoli sia nei titoli stessi. Consultando i prospetti IPO, gli investitori possono trovare informazioni finanziarie verificate negli ultimi tre anni finanziari, insieme a un rapporto di revisione per ogni anno.

L'EGM consente anche la quotazione di aziende più piccole con un requisito del 10% di quota libera. Le aziende che desiderano quotare le loro azioni sull'EGM devono solo consegnare a Borsa Italiana un Documento di Ammissione, che è meno dettagliato del prospetto sopra menzionato. In questo mercato, la due diligence non è condotta dalle autorità italiane ma da un consulente nominato (NOMAD) che è responsabile della verifica del rispetto dei requisiti.

In termini di requisiti di divulgazione dopo la quotazione, le aziende quotate su EXM e EGM devono divulgare i loro bilanci annuali e i rapporti semestrali al mercato. La Commissione Nazionale per le Società e la Borsa (CONSOB) e Borsa Italiana S.p.A. hanno ruoli fondamentali nel mercato finanziario italiano. Borsa Italiana gestisce il mercato azionario italiano, definendo le procedure che le aziende quotate devono seguire per accedere al mercato e approvando o rifiutando le richie-

ste di quotazione in borsa. CONSOB è un'autorità di vigilanza che ha il compito di proteggere gli investitori e garantire la trasparenza e l'efficienza dei mercati finanziari italiani.
Per quanto riguarda il mercato delle IPO, CONSOB è inizialmente responsabile dell'approvazione dei prospetti IPO. Successivamente, supervisiona l'intero processo di IPO, garantendo che si svolga in conformità con le norme e i regolamenti italiani e comunitari. Tipicamente, il processo di IPO italiano sull'EXM richiede da cinque a sette mesi, mentre il processo di quotazione sull'EGM è semplificato e di solito richiede da tre a quattro mesi per essere completato.

1.2 L'Ipotesi di Efficienza del Mercato e l'underpricing delle IPO

La letteratura descrive l'underpricing delle IPO come la differenza tra il prezzo di chiusura nel giorno di quotazione e il prezzo di offerta. Tuttavia, il processo di IPO coinvolge due mercati distinti, ovvero il mercato primario e secondario: il mercato primario per l'apertura e il mercato secondario per la chiusura, con caratteristiche fondamentalmente diverse. Il mercato primario infatti è il luogo dove vengono emessi nuovi titoli finanziari per la prima volta, mentre il mercato secondario è il luogo dove i titoli già esistenti vengono scambiati tra gli investitori dopo la loro emissione iniziale.
Risulta di interesse andare a comprendere le dinamiche di mercato che contraddistinguono il processo di quotazione in borsa; un prezzo troppo elevato infatti può portare al fallimento della IPO mentre un prezzo troppo basso trasferisce ricchezza dai vecchi ai nuovi azionisti. In accordo con l'ipotesi d'efficienza dei mercati, in un mercato efficiente i prezzi delle azioni dovrebbero riflettere tutte le informazio-

ni disponibili. Tuttavia, diversi tipi di informazioni possono influenzare i valori di sicurezza del mercato. A tal proposito, Clarke, et al. (2001) definiscono tre livelli di efficienza di mercato, che si distinguono per la quantità e il tipo di informazioni riflesse nei prezzi dei titoli: forma debole, forma semi-forte e forma forte:

- La forma debole dell'Ipotesi d'Efficienza del Mercato asserisce che il prezzo corrente incorpori pienamente solo le informazioni contenute nella storia passata dei prezzi. Cioè, nessuno può rilevare titoli con un prezzo errato e battere il mercato analizzando i prezzi passati.

- La forma semi-forte dell'Ipotesi d'Efficienza del Mercato suggerisce che il prezzo corrente incorpori completamente tutte le informazioni disponibili al pubblico.

- La forma forte dell'Ipotesi d'Efficienza del Mercato afferma che il prezzo corrente incorpora completamente tutte le informazioni esistenti, sia pubbliche che private. La principale differenza tra l'ipotesi di efficienza semi-forte e quella forte è che in quest'ultimo caso nessuno dovrebbe essere in grado di generare profitti sistematicamente, anche se si fa trading su informazioni non note al pubblico al momento.

La letteratura sul mercato azionario italiano, come discusso da Cervellati et al. (2008), identifica una forma di efficienza semi-forte nel mercato azionario italiano. In questa pubblicazione viene evidenziato come le modifiche nelle raccomandazioni degli analisti finanziari impattino significativamente i prezzi delle azioni, sia in termini di rendimenti che di volumi di scambio. Queste reazioni si osservano non solo dopo la pubblicazione delle informazioni al pubblico, ma anche prima. Ciò indica che le informazioni pubblicamente disponibili non sono immediatamente incorporate nei prezzi delle azioni, suggerendo l'esistenza di inefficienze nel mercato.

Nonostante la presenza di un'efficienza semi-forte, dove teoricamente tutte le informazioni pubbliche dovrebbero essere già incorporate nei prezzi, l'underpricing persiste come strategia per mitigare specifici rischi e attrarre investitori, suggerendo così delle inefficienze pratiche nel mercato. In letteratura esistono una serie di teorie per cui l'underpricng viene accettato sistematicamente dagli azionisti:

- Modelli di informazione asimmetrica: Come esemplificato da Jegadeesh et al. (1993), l'approccio classico delle teorie dell'informazione asimmetrica suggerisce che gli emittenti possano essere più informati degli investitori riguardo al vero valore delle aziende che diventano pubbliche. In alternativa, se gli investitori sono più informati rispetto all'emittente, ad esempio riguardo alla domanda generale di mercato per le azioni, l'emittente si trova di fronte a un problema di collocazione, non conoscendo il prezzo che il mercato è disposto a sostenere.

- Teoria delle Prospettive: Loughran e Ritter (2002) utilizzano la teoria delle prospettive di Kahneman e Tversky (1979) per spiegare che gli imprenditori tendono ad accettare un eccessivo underpricing se apprendono contemporaneamente di una valutazione di mercato post-IPO superiore alle aspettative iniziali. Questo fenomeno si verifica perché, secondo la teoria delle prospettive, gli individui valutano le decisioni basandosi sui guadagni e perdite relativi a un punto di riferimento. Un significativo aumento della ricchezza post-IPO riduce la percezione della perdita associata all'underpricing, portando gli imprenditori a negoziare meno intensamente il prezzo di offerta con i sottoscrittori.

- Effetto di Signalling: La teoria della segnalazione afferma che le aziende possono utilizzare l'underpricing delle IPO come un segnale positivo per gli investitori. Secondo questa teoria infatti, offrendo le azioni a un prezzo inferiore

al loro valore intrinseco, le aziende possono creare una buona impressione agli investitori, costruendo così una reputazione positiva. Questo approccio dovrebbe permettere alle aziende di effettuare future emissioni di azioni a prezzi più alti, beneficiando della fiducia guadagnata inizialmente. (Ritter e Welch 2002)

- Controversie legali: Alcuni studiosi, come Hughes e Thakor (1992), sostengono che la pratica dell'underpricing sia utilizzata per ridurre il rischio di cause legali, dato che un prezzo iniziale più basso riduce la probabilità di una causa se il prezzo delle azioni dovesse poi scendere. Tuttavia, altre ricerche indicano che l'underpricing non necessariamente protegge dalle cause legali e potrebbe essere dovuto al fatto che le IPO più rischiose o controverse sono naturalmente più inclini a essere citate in giudizio. Ad esempio, uno studio condotto in Giappone da Beller et al. (1992) supporta questa tesi. Inoltre, la prova più convincente che la responsabilità legale non è il principale fattore che determina l'underpricing è data dal fatto che i paesi senza le stesse tendenze legali degli Stati Uniti presentano livelli simili di underpricing (Keloharju (1993)).

La scelta del prezzo delle IPO, come illustrato dai meccanismi precedentemente descritti rappresenta un complesso equilibrio tra attrarre investitori e riflettere il valore dell'azienda. Un prezzo troppo alto rischia di allontanare gli investitori più esperti, mentre un prezzo troppo basso potrebbe suggerire una mancanza di fiducia nell'azienda portando alla perdita di capitale per gli azionisti originali. Dunque, comprendere i diversi livelli di informazione posseduti dagli investitori e le loro potenziali reazioni ai prezzi delle IPO risulta fondamentale per il successo di queste operazioni.

1.3 L'Influenza dei media e dei social media sull'underpricing delle IPO

La letteratura che studia il ruolo dei media nella finanza è in continua evoluzione. Gli studiosi utilizzano dati provenienti da diverse fonti mediatiche, come articoli di giornale e social media, per analizzare il loro impatto su questioni come i rendimenti di mercato, la liquidità e l'underpricing delle IPO. Esiste una letteratura robusta che utilizza i dati dei social media per prevedere i movimenti del mercato azionario. Ad esempio, studi come quello di Sprenger et al. (2014) hanno dimostrato che i tweet possono prevedere i movimenti dei prezzi delle azioni e che i sentimenti espressi in tali tweet hanno una relazione significativa con i rendimenti, la liquidità e i volumi. Questa ricerca sottolinea l'importanza crescente delle analisi basate sui media e dei sentimenti espressi sui social, e di come questi possano fornire indicazioni sulle tendenze del mercato e sul comportamento degli investitori.

In sintesi, l'impatto del sentiment sui mercati finanziari, parte della finanza comportamentale, sta rapidamente diventando un elemento centrale della finanza tradizionale (Ritter, 2003). Lundmark et al. (2017) hanno osservato che le variabili riferite ai social media influenzano il livello di underpricing di un'IPO. Lo studio ha analizzato diverse variabili come l'esistenza di un account Twitter, il numero di follower, il numero di tweet e retweet, e altre metriche per spiegare i fenomeni di underpricing. I risultati hanno mostrato che avere un account Twitter, utilizzare efficacemente tale account e l'alto numero di tweet e retweet sono associati a un livello di underpricing più elevato.

Tuttavia, la carenza di studi sui social media nel caso italiano rende difficile comprendere direttamente il loro impatto sull'underpricing delle IPO. Ciò potrebbe essere dovuto al non coinvolgimento degli investitori italiani sui social media per discutere e informarsi di questioni tecniche. Invece, gli articoli dei media possono

essere utilizzati per generare fiducia negli investitori. Come discusso nel capitolo 1.2, l'impatto delle informazioni sulle performance delle IPO è cruciale per comprendere le dinamiche di underpricing nel mercato azionario italiano.
Non sempre la quantità totale di informazioni risulta significativa per l'underpricing; piuttosto, è la qualità di tali informazioni ad avere un impatto determinante. Ad esempio, il sostegno da parte di una rinomata società di venture capital può fungere da segnale indiretto della qualità di un'IPO, come evidenziato da Chua (2014). I media, inoltre, comunicano direttamente la qualità delle IPO. Gupta et al. (2022) hanno individuato una relazione positiva tra il sentiment mediatico e l'underpricing; tuttavia, il loro studio dimostra che il numero di articoli pubblicati non svolge un ruolo significativo nello spiegare l'underpricing delle IPO, in contrasto con i risultati di Liu et al. (2014a, 2014b). La letteratura porta fermamente a credere che un punteggio di sentiment positivo e una maggiore copertura mediatica, misurata dal numero di articoli, aumenti la fiducia degli investitori nell'IPO, causando un maggiore underpricing delle IPO. Pertanto, ipotizziamo quanto segue:

- L'underpricing delle IPO è positivamente associato a variabili legate ai social media.
- L'underpricing delle IPO è positivamente associato al sentiment dei media.
- L'underpricing delle IPO è positivamente associato al numero di articoli dei media.

Nei capitoli successivi verrà esteso il concetto di sentiment, che non si limiterà alla sola valutazione positiva e negativa dei contenuti, ma includerà anche l'analisi dell'incertezza espressa nei post sui social media e negli articoli dei media finanziari. Questo approccio consentirà di ottenere una comprensione più completa delle percezioni del mercato e di come queste influenzino l'underpricing delle IPO.

Capitolo 2

Raccolta dei dati

Il presente capitolo è dedicato alla descrizione del processo di raccolta dei dati impiegati per studiare l'impatto delle dinamiche mediatiche e delle interazioni sui social network sull'underpricing delle IPO italiane.
Un fondamentale contributo a questa ricerca è stato fornito dal professor Riccardo Ferretti dell'Università di Modena e Reggio Emilia, che ha messo a disposizione la lista delle 243 aziende quotate in borsa dal 2010 al 2022 e le variabili economico-finanziarie utilizzate per l'analisi.
Queste variabili, fondamentali per comprendere i fenomeni finanziari associati alle IPO, saranno descritte nel capitolo 3. Per analizzare i dati ottenuti dai social media e dai media tradizionali, è stato utilizzato il linguaggio di programmazione R, al fine di automatizzare le fasi di estrazione e analisi dei dati, quali post di Facebook, dati da Google Trends e articoli finanziari dai siti di SoldiOnline e MilanoFinanza.

2.1 CrowdTangle

Per la raccolta dei dati relativi ai post su Facebook, è stato creato manualmente un elenco di pagine Facebook delle aziende coinvolte nelle IPO, utilizzando la

piattaforma CrowdTangle. Successivamente, è stato utilizzato un approccio automatizzato che ha permesso di estrarre informazioni dalle pagine Facebook delle aziende, sfruttando l'API di CrowdTangle. Questo metodo ha consentito di focalizzare l'attenzione sui dati specifici delle aziende di interesse e di creare un archivio social di queste.

Le API (Application Programming Interfaces) sono un insieme di definizioni e protocolli che permettono a un software di comunicare con un altro. Esse facilitano l'integrazione di funzioni e servizi tra diverse applicazioni e sistemi, consentendo lo scambio di dati e l'esecuzione di operazioni in modo automatizzato. Le API sono fondamentali nella tecnologia moderna poiché permettono l'interpretabilità tra diversi sistemi e applicazioni, facilitando lo sviluppo di software complessi.

L'utilizzo dell'API di CrowdTangle ha permesso di automatizzare il processo di raccolta dei dati relativi ai post su Facebook. Il processo è iniziato con la preparazione di un dataset contenente le informazioni necessarie per ciascuna azienda, inclusi l'identificativo della pagina Facebook e le date rilevanti, quali due settimane prima della quotazione e la data di fine periodo di osservazione. Per ogni azienda, le date di inizio e fine periodo sono state formattate secondo il formato richiesto dall'API. Successivamente, utilizzando un ciclo iterativo, sono state inviate richieste GET all'API di CrowdTangle per ogni azienda, specificando il periodo di interesse che copriva le due settimane precedenti la quotazione. La richiesta includeva parametri come il token di accesso, l'identificativo della pagina Facebook e i criteri di ordinamento.

Le risposte ricevute dall'API sono state verificate per garantirne il successo e, in caso di esito positivo, i dati dei post sono stati estratti e salvati in una lista, associandoli all'azienda corrispondente. Nei casi di errore, come l'inesistenza o l'inaccessibilità della pagina, è stato registrato un messaggio di errore e l'azienda è stata contrassegnata come non recuperata.

I post recuperati sono stati poi filtrati escludendo i post privi di contenuto testuale. Per ciascuna azienda, è stata calcolata la somma delle interazioni per i post filtrati, fornendo una visione complessiva dell'engagement sui social media che potrebbe influenzare la percezione pubblica dell'IPO.
I dati relativi alle interazioni sui post sono stati successivamente aggiunti al dataset principale. Nei casi in cui non erano presenti post per una determinata azienda, il valore è stato posto pari a zero per mantenere la coerenza del dataset. Inoltre, è stata introdotta una variabile binaria per identificare l'esistenza della pagina Facebook al momento della quotazione in borsa.
Questo approccio ha permesso di ottenere un archivio completo dei post Facebook, permettendo l'analisi dell'impatto delle attività sui social media.

2.2 Google Trends

Per la raccolta dei dati relativi alle tendenze di ricerca su Google, è stato utilizzato un approccio automatizzato che ha sfruttato l'API di Google Trends attraverso il pacchetto `gtrendsR` in R. Questo metodo ha consentito di ottenere informazioni sulle tendenze di ricerca per le aziende coinvolte nelle IPO. Si è scelto di analizzare l'interesse di ricerca per ciascuna azienda nel periodo che va dal mese precedente alla quotazione in borsa fino al giorno prima della quotazione stessa, per valutare l'evoluzione dell'attenzione pubblica in relazione agli eventi di IPO.
Il processo di estrazione dei dati è stato automatizzato tramite l'uso dell'API di Google Trends. Per ciascuna azienda presente nel dataset, è stata formulata una query di ricerca specifica per il nome dell'azienda, limitando la ricerca al mese predente la quotazione in borsa. La funzione gtrends è stata utilizzata per recuperare i dati di tendenza, che includono il numero di ricerche (hits) per ciascuna data, consentendo un'analisi delle fluttuazioni dell'interesse nel tempo.

A causa delle limitazioni imposte da Google sull'uso frequente del suo servizio, è stato necessario implementare pause programmate tra le richieste per evitare problemi di rate limiting.
Inoltre, per ciascuna azienda, è stata creata una variabile binaria indicante se ci sia stato un picco di ricerche (il massimo di hits) nei quindici giorni precedenti la data di quotazione. Questo è stato fatto per identificare un potenziale aumento dell'interesse del pubblico o dell'attività promozionale poco prima dell'IPO.
Alla fine di questo processo è stata quindi aggiunta una nuova variabile al dataset che riflette il picco di interesse di ricerca. L'analisi di questi dati aiuta a comprendere se l'attenzione del pubblico possa avere una correlazione con l'underpricing osservato nelle IPO.

2.3 SoldiOnline

Per quanto riguarda l'analisi dei contenuti mediatici da fonti online, è stata utilizzata una strategia di web scraping per estrarre articoli dal sito "SoldiOnline". Questo sito è una delle principali risorse finanziarie italiane, e rappresenta una fonte di notizie e analisi per le aziende che si apprestano a quotarsi in borsa.
Il processo di estrazione degli articoli è stato automatizzato attraverso una funzione personalizzata in R, che accetta come parametri il nome dell'azienda, la data di inizio e la data di fine del periodo di interesse. Il periodo di interesse è stato definito come le due settimane precedenti alla data di quotazione in borsa dell'azienda, per comprendere l'attenzione mediatica nel periodo antecedente all'IPO.
Il processo di scraping ha coinvolto la navigazione delle pagine di ricerca del sito SoldiOnline, utilizzando come chiavi di ricerca i nomi delle aziende. Per ciascuna pagina, il codice estrae le date di pubblicazione degli articoli, i titoli e i link agli articoli stessi. Solo gli articoli pubblicati nel periodo di interesse sono stati conser-

vati per l'analisi successiva.
Per facilitare il processo di estrazione dei dati dal sito web di SoldiOnline, è stata utilizzata l'estensione per Google Chrome chiamata SelectorGadget. Questo strumento ha permesso di identificare i selettori CSS necessari per isolare le componenti specifiche delle pagine web, come le date di pubblicazione, i titoli e i link agli articoli.
Dopo il recupero degli articoli, è stata implementata un'altra funzione, per scaricare il contenuto testuale di ciascun articolo. Questi testi sono stati poi analizzati per estrarre le informazioni che verranno presentate nel prossimo capitolo.

2.4 MilanoFinanza

Per quanto riguarda l'analisi dei contenuti mediatici da fonti online, è stata utilizzata una strategia di web scraping per estrarre articoli dal sito "MilanoFinanza". MilanoFinanza è un portale di informazioni finanziarie, che fornisce analisi e aggiornamenti sugli sviluppi del mercato.
Il processo di estrazione degli articoli è stato automatizzato attraverso script in linguaggio R, che hanno permesso di effettuare il login al sito per accedere alle aree riservate e di ricercare gli articoli di interesse. Questa ricerca è stata effettuata inserendo il nome dell'azienda nella barra di ricerca e impostando il periodo di due settimane prima della data di quotazione come intervallo temporale di interesse.
Durante la navigazione delle pagine di risultato di MilanoFinanza, il codice ha estratto le date di pubblicazione, i titoli e i link agli articoli. L'accesso ai contenuti degli articoli è stato poi ottenuto seguendo i link raccolti, e i contenuti testuali sono stati scaricati tramite l'utilizzo di SelectorGadget. Questi contenuti sono stati analizzati per estrarre informazioni rilevanti che verranno presentate nel capitolo successivo.

Capitolo 3

Descrizione del dataset e analisi esplorativa

In questo capitolo, verrà fornita una descrizione completa del dataset utilizzato nell'analisi, suddividendo le variabili in categorie economiche e derivate dai media e social media. Verrà inizialmente introdotta la variabile risposta, ovvero il rendimento di mercato al primo giorno. Inoltre, verrà presentata anche una prima analisi esplorativa per comprendere meglio la distribuzione e le caratteristiche principali delle variabili. I dati relativi alle variabili economiche sono stati forniti dal Prof. Ferretti, come citato all'inizio del Capitolo 2. Dal capitolo 3.3 verranno descritte le variabili derivate dall'analisi dei dati raccolti dai media e dai social media.

3.1 Descrizione della variabile risposta

In questa sezione, verrà descritta la variabile risposta, il rendimento netto dell'IPO nel primo giorno di quotazione, che rappresenta l'oggetto dello studio.
Il rendimento netto dell'IPO nel primo giorno di quotazione rappresenta la variazione del prezzo delle azioni di un'azienda dal prezzo di offerta iniziale al prezzo di chiusura del primo giorno di negoziazione. La formula utilizzata per calcolare il rendimento netto dell'IPO nel primo giorno è la seguente:

$$y = \left(\frac{\text{Prezzo di Chiusura del Primo Giorno} - \text{Prezzo di Offerta}}{\text{Prezzo di Offerta}} \right) \tag{3.1}$$

3.1.1 Analisi esplorativa della variabile risposta

La figura 3.1 mostra l'istogramma del rendimento netto dell'IPO nel primo giorno di quotazione.

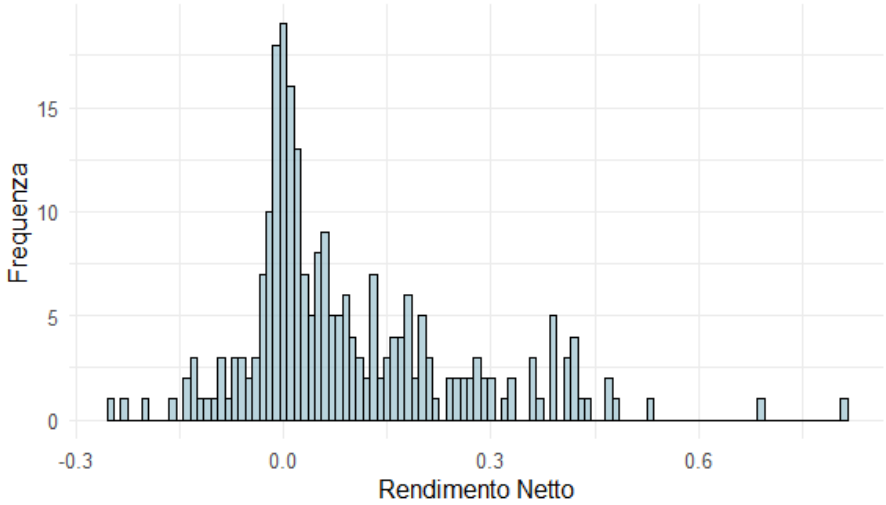

Figura 3.1: Istogramma del Rendimento Netto dell'IPO nel Primo Giorno

Dal grafico 3.1 è possibile notare che la densità si concentra intorno allo zero,

con una leggera asimmetria verso destra. Questo suggerisce che la maggior parte dei rendimenti netti si trova vicino allo zero, ma ci sono alcuni valori estremi che manifestano la presenza di rendimenti netti positivi elevati.
Le statistiche descrittive supportano questa osservazione. Il rendimento netto minimo è -0.247784, mentre il massimo raggiunge 0.808411. La mediana del rendimento netto è 0.044577, indicando che il 50% delle IPO ha un rendimento netto inferiore a questo valore. La media del rendimento netto è 0.093142, superiore alla mediana, suggerendo una leggera asimmetria positiva nella distribuzione.
Analizzando i quartili, il primo quartile è -0.005087, indicando che il 25% delle IPO ha avuto un rendimento netto negativo nel primo giorno. Il terzo quartile è 0.169507, suggerendo che il 75% delle IPO ha un rendimento netto inferiore a questo valore.
La presenza di alcuni rendimenti netti molto elevati, con un massimo di 0.808411, indica la presenza di outlier positivi. Questi outlier possono influenzare la media e la distribuzione complessiva dei dati. Tuttavia, sono indicativi di un fenomeno comune nel mercato delle IPO, dove alcune emissioni in borsa registrano guadagni molto elevati nel primo giorno di quotazione. Questo comportamento riflette l'underpricing, che può portare a rendimenti iniziali elevati per alcune IPO.
In sintesi, l'analisi dell'istogramma del rendimento netto delle IPO nel primo giorno di quotazione mostra una distribuzione prevalentemente concentrata intorno allo zero, con una leggera tendenza positiva. La presenza di outlier positivi conferma che, sebbene la maggior parte delle IPO registri rendimenti moderati, alcune possono ottenere guadagni molto elevati.
In conclusione, è stato effettuato il test di Kolmogorov-Smirnov, il quale ha mostrato che la variabile risposta non segue una distribuzione normale. Quindi, nelle sezioni successive, per valutare l'impatto delle variabili esplicative sulla risposta, verranno effettuati i test non parametrici di Spearman, Mann-Whitney e Kruskal-Wallis,

data la natura non lineare della risposta.

3.2 Descrizione delle variabili economiche

In questa sezione verranno descritte le variabili economiche incluse nel dataset.
La prima variabile economica è MKT, che rappresenta il mercato di quotazione dell'IPO. Questa variabile è dicotomica e assume un valore pari a 1 se l'IPO è quotata su EGM (Euronext Growth Milan), e 0 se quotata su EXM (Euronext Milan).
Il trend di mercato, indicato come market_index, è misurato dal rendimento dell'indice FTSE Italia All-Share nei 100 giorni precedenti la quotazione dell'azione. Questo valore fornisce una misura del trend generale del mercato azionario italiano nel periodo precedente l'IPO.
La volatilità di mercato, indicata come market_volatility, è misurata dalla deviazione standard del rendimento dell'indice FTSE Italia All-Share nei 60 giorni precedenti la quotazione dell'azione. Questo parametro riflette la stabilità del mercato nel periodo precedente l'IPO.
La quota di capitale collocata nell'IPO, rappresentata come placed_stake, è calcolata come il numero di azioni collocate sul mercato diviso il totale delle azioni esistenti. Questo valore indica la percentuale del capitale sociale che viene offerta al pubblico durante l'IPO.
La variabile "sponsor" rappresenta la banca o il soggetto finanziario che sponsorizza l'IPO. Una società che intenda chiedere la quotazione in borsa deve nominare uno sponsor, ossia un soggetto che collabora nell'assolvimento degli impegni connessi con l'ammissione alla quotazione e svolge funzioni di garante in merito al profilo qualitativo dell'emittente stesso. La presenza di uno sponsor potrebbe influenzare significativamente il successo dell'IPO, poiché lo sponsor fornisce credibilità e supporto tecnico durante il processo di quotazione.

La variabile log_cap_raised rappresenta il logaritmo del capitale raccolto dall'IPO. Si è scelto di utilizzare il logaritmo naturale poichè aiuta a stabilizzare la variabilità presente nei dati.
La variabile capital_raised rappresenta la quota del capitale raccolto tramite azioni di nuova emissione nell'IPO. Questa misura indica quanto del capitale raccolto proviene dalla creazione di nuove azioni.
Infine, la variabile Tech è una variabile dicotomica che assume il valore di 1 se un'azienda appartiene al settore Technology e 0 altrimenti. Questa variabile permette di identificare se l'azienda opera nel settore tecnologico, che può avere dinamiche di mercato diverse rispetto ad altri settori.

3.2.1 Analisi esplorativa delle variabili economiche quantitative

In questa sezione vengono analizzate le statistiche descrittive delle variabili economiche quantitative considerate nello studio. La Tabella 3.1 riassume le principali caratteristiche di queste variabili, quali i valori di minimo, primo quartile (Q1), mediana, media, terzo quartile (Q3) e massimo.

Tabella 3.1: Statistiche descrittive delle variabili economiche quantitative

Variabile	Minimo	Q1	Mediana	Media	Q3	Massimo
capital_raised	0.0000	0.7541	1.0000	0.8051	1.0000	1.0002
log_cap_raised	-1.0969	0.5628	0.8376	1.0472	1.3549	3.5268
market_index	-0.3329	-0.0201	0.0550	0.0409	0.1023	0.2689
placed_stake	0.0112	0.1795	0.2500	0.2700	0.3220	1.0000
market_volatility	0.0060	0.0095	0.0112	0.0121	0.0135	0.0319

La Tabella 3.1 presenta le statistiche descrittive delle variabili economiche quantitative considerate nello studio. La variabile capital_raised mostra un valore mediano pari a 1 e una media pari a 0.8051, suggerendo che nella maggior parte dei

casi il capitale raccolto è vicino all'intero ammontare delle azioni emesse. Tuttavia, il valore minimo di 0 indica che ci sono casi in cui non è stato raccolto capitale.
Per quanto riguarda la variabile log_cap_raised, essa presenta un'ampia gamma di valori, con un minimo di -1.0969 e un massimo di 3.5268. La mediana e il primo quartile suggeriscono che la maggior parte dei valori sono concentrati nella fascia inferiore della distribuzione.
Il trend di mercato, rappresentato dalla variabile market_index, ha una media positiva di 0.0409, indicando che, mediamente, l'indice FTSE Italia All-Share ha avuto una performance positiva nei 100 giorni precedenti l'IPO. Tuttavia, la presenza di un minimo negativo di -0.3329 evidenzia che ci sono stati periodi di declino del mercato.
La quota di capitale collocata, rappresentata dalla variabile placed_stake, varia notevolmente, con un minimo di 0.0112 e un massimo di 1.0000. La media di 0.2700 suggerisce che, in media, il 27% del capitale sociale è stato offerto durante l'IPO.
Infine, la variabile market_volatility ha una mediana di 0.0112, con un primo quartile di 0.0095 e un terzo quartile di 0.0135. Questo indica che vi è una variazione relativamente ridotta nella volatilità del mercato nei 60 giorni precedenti la quotazione.

3.2.2 Analisi esplorativa delle variabili qualitative categoriali

In questa sezione vengono analizzate le variabili economiche qualitative categoriali presenti nel dataset. Le tabelle seguenti riassumono le frequenze relative delle variabili categoriali considerate nello studio.

Tabella 3.2: Frequenze relative della variabile MKT

Categoria	Frequenza Relativa
0	0.1893
1	0.8107

La Tabella 3.2 mostra che l'81.07% delle IPO sono quotate su EGM (Euronext Growth Milan), mentre il restante 18.93% su EXM (Euronext Milan). Questo indica una predominanza delle quotazioni su EGM nel campione analizzato.

Tabella 3.3: Frequenze relative della variabile Sponsor

Categoria	Frequenza Relativa
Advance SIM	0.0370
Alantra Capital Markets	0.0412
Altri sponsor	0.1029
Banca Finnat Euramerica	0.0741
Banca IMI S.p.A.	0.0494
Banca Popolare di Vicenza	0.0288
Banca Profilo	0.0247
BPER Banca	0.0370
CFO SIM S.p.A.	0.0247
EnVent Capital Markets	0.1646
Equita SIM S.p.A.	0.0494
Integrae SIM S.p.A.	0.2428
Intermonte SIM	0.0206
Mediobanca	0.0823
Unicredit CIB	0.0206

La Tabella 3.3 presenta le frequenze relative delle varie categorie della variabile Sponsor, che rappresenta la banca o il soggetto finanziario che sponsorizza l'IPO. Integrae SIM S.p.A. emerge come lo sponsor più frequente, rappresentando il 24.28% del totale delle IPO, seguito da EnVent Capital Markets, che sponsorizza il 16.46% delle IPO. Altri sponsor come Mediobanca e Banca Finnat Euramerica mostrano una presenza significativa, rispettivamente con una quota dell'8.23% e del

7.41%. La categoria "Altri sponsor" include sponsor che hanno partecipato a meno di 5 IPO, rappresentando il 10.29% del totale. Questa categoria è stata creata per aggregare gli sponsor meno frequenti e fornire una visione più chiara degli sponsor maggiormente attivi.

Tabella 3.4: Frequenze relative della variabile Technology

Categoria	Frequenza Relativa
0	0.8354
1	0.1646

La Tabella 3.4 evidenzia che il 83.54% delle aziende non appartiene al settore tecnologico, mentre il 16.46% sì. Questo suggerisce una predominanza di aziende non tecnologiche nel campione.

3.2.3 Relazioni tra l'underpricing e le variabili quantitative

La Figura 3.2 mostra gli scatter plot tra il livello di underpricing (y) e le cinque variabili economiche quantitative: market_index, capital_raised, log_cap_raised, placed_stake e market_volatility.
Dall'analisi dei grafici, emergono i seguenti punti salienti:

- Market Index vs Underpricing (y): Il grafico scatter non mostra una relazione tra il trend di mercato e il livello di underpricing. I dati sono distribuiti in modo uniforme lungo l'asse delle ordinate, suggerendo che il trend di mercato potrebbe non avere un'influenza significativa sull'underpricing.

- Capital Raised vs Underpricing (y): Anche in questo caso, il grafico scatter non indica una chiara relazione tra il capitale raccolto e l'underpricing. La distribuzione dei punti sembra essere piuttosto sparsa, senza una direzione chiara, suggerendo una relazione debole o inesistente tra queste due variabili.

- Log Capital Raised vs Underpricing (y): Il grafico scatter per il logaritmo del capitale raccolto mostra una dispersione simile a quella del capitale raccolto, senza un pattern evidente. Questo implica che la trasformazione logaritmica del capitale raccolto non evidenzia una relazione chiara con l'underpricing.

- Placed Stake vs Underpricing (y): Il grafico scatter indica una leggera concentrazione di punti per valori bassi di placed stake, ma la distribuzione rimane abbastanza diffusa. Non si osserva un trend chiaro, suggerendo che l'influenza della quota di capitale collocata sull'underpricing potrebbe non essere significativa.

- Market Volatility vs Underpricing (y): Infine, il grafico scatter per la volatilità di mercato mostra una distribuzione relativamente uniforme dei punti, senza indicazioni di una relazione diretta con l'underpricing. Anche in questo caso, non sembra esserci una correlazione forte tra la volatilità del mercato e l'underpricing.

Figura 3.2: Scatter plot tra underpricing (y) e variabili economiche quantitative

In conclusione, data la distrbuzione non normale della variabile risposta, è stata effettuata l'analisi della correlazione di Spearman tra la risposta e le variabile economiche quantitative, questi hanno mostrato l'assenza di correlazione, confermando l'analisi degli scatter plot.

3.2.4 Relazioni tra l'underpricing e le variabili qualitative

La Figura 3.3 mostra i box plot tra il livello di underpricing (y) e i livelli delle variabili economiche qualitative MKT e Technology.
Dall'analisi dei grafici, emergono i seguenti aspetti rilevanti:

- MKT vs Underpricing (y): Il box plot mostra che la distribuzione di y varia tra i due livelli di MKT. In particolare, le IPO quotate su EGM (MKT = 1) tendono ad avere un underpricing maggiore rispetto a quelle quotate su EXM, suggerendo che il mercato di quotazione potrebbe influenzare l'underpricing.

- Technology vs Underpricing (y): Il box plot per la variabile Technology indica che le aziende che appartengono al settore tecnologico (Technology = 1) tendono ad avere un underpricing maggiore rispetto alle aziende che non appartengono a questo settore. La mediana dell'underpricing è più alta per le aziende Tech, suggerendo che il settore di appartenenza potrebbe influenzare il livello di underpricing.

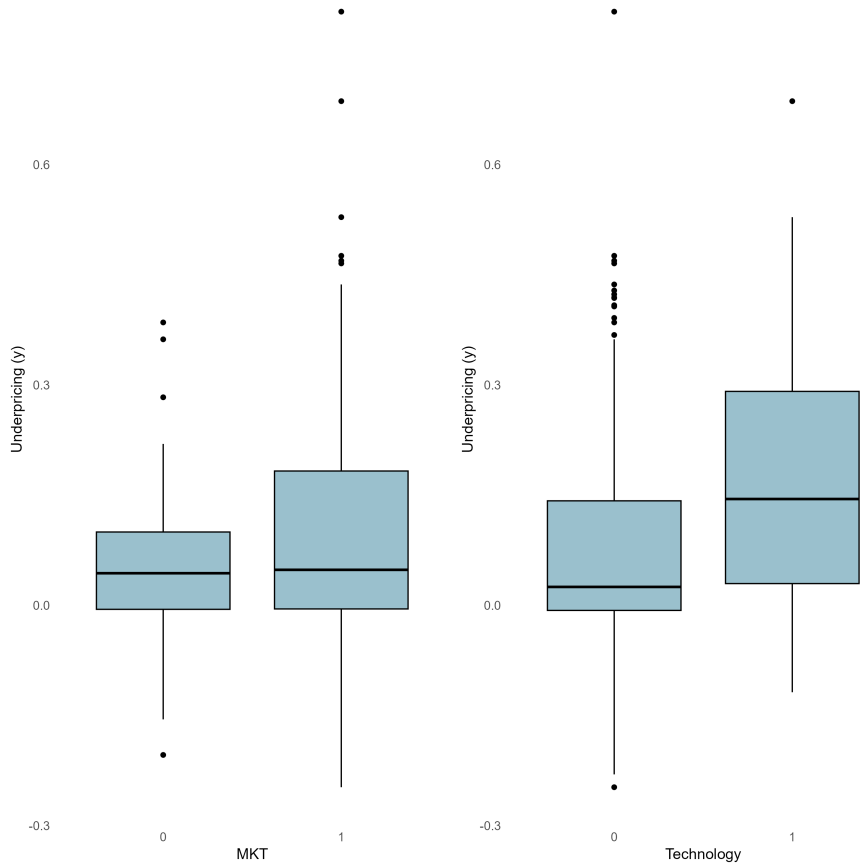

Figura 3.3: Box plot tra underpricing (y) e le variabili qualitative MKT e Technology IPO

La Figura 3.4 mostra il box plot tra il livello di underpricing (y) e ciascuno sponsor.

Dall'analisi del grafico, emergono i seguenti aspetti rilevanti: la distribuzione di y varia tra i diversi sponsor, nello specifico alcuni sponsor, come Banca Profilo e BPER Banca, presentano una maggiore variabilità e livelli di underpricing più elevati rispetto ad altri sponsor; altri sponsor, come Banca Popolare di Vicenza e Unicredit CIB, mostrano livelli di underpricing più bassi e una minore variabilità.

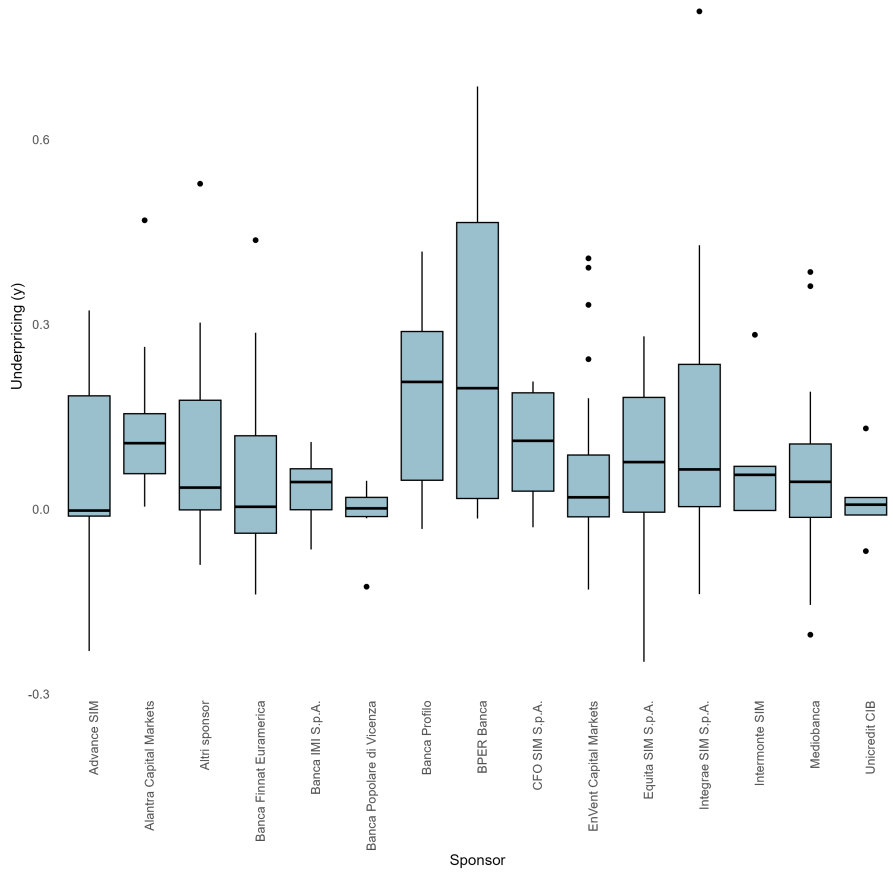

Figura 3.4: Box plot tra underpricing (y) e la variabile qualitativa Sponsor

In conclusione, sono stati effettuati i test di Wilcoxon-Mann-Whitney per le 2 variabili MKT e Technology, e il test di Kruskal-Wallis per la variabile sponsor. L'analisi dei test ha confermato i risultati dell'analisi grafica per la variabile Technology, il test ha fornito una statistica pari a -3.6177 e p-value = 0.0002972. La

variabile MKT invece, non ha mostrato differenze significative tra i due gruppi. Per quanto riguarda la variabile sponsor, il test ha mostrato l'assenza di differenze tra gruppi ad un livello di significatività dell'5%, ottenendo tuttavia un p-value pari a 0.09653.

3.3 Variabili derivate da Facebook

Le variabili derivate da Facebook includono "In_Fb_List", una variabile dicotomica che indica la presenza di una pagina Facebook per l'azienda. Un valore pari a 1 indica che l'azienda possiede una pagina Facebook, mentre un valore pari a 0 indica l'assenza di tale pagina. Questa variabile è utile per identificare se l'azienda utilizza Facebook come canale di comunicazione ufficiale.

La variabile "Interazioni_totali" rappresenta il numero complessivo di interazioni sui post aziendali nelle due settimane precedenti l'IPO. Questa misura include diverse forme di interazione che riflettono il coinvolgimento degli utenti con i contenuti pubblicati dall'azienda sui social media, offrendo un indicatore dell'efficacia della loro strategia di comunicazione. Per quanto riguarda i post su Facebook, le interazioni totali comprendono le reazioni degli utenti ai post, che possono essere espresse tramite varie emoji come "like", "heart", "sad", "angry", "haha", "wow" e "care". Le interazioni includono anche il numero di condivisioni dei post, indicando la diffusione dei contenuti, e il numero di commenti ricevuti sui post, che riflettono l'engagement diretto degli utenti con i contenuti. Questa misura fornisce un'indicazione del livello di attività e del coinvolgimento del pubblico con i contenuti dell'azienda.

Infine, "facebook_ipo" è una variabile dicotomica che assume un valore pari a 1 se l'azienda ha menzionato l'IPO nei propri post su Facebook nelle due settimane antecedenti la quotazione, e 0 in caso contrario. Questa variabile consente di valutare

se l'azienda ha utilizzato Facebook per promuovere l'IPO. Per le 59 aziende per cui è stato possibile scaricare i dati, sono stati esaminati un totale di 573 post pubblicati su Facebook. Ogni post è stato analizzato per verificare la presenza di menzioni relative all'IPO. In base a questa analisi, è stato assegnato il valore 1 se almeno un post dell'azienda menzionava l'IPO nel periodo considerato, altrimenti è stato assegnato il valore 0. Questo processo ha permesso di determinare con precisione se l'azienda ha sfruttato la piattaforma di Facebook come strumento di comunicazione per la propria offerta pubblica iniziale.

È importante notare che i post sono stati scaricati solo per 59 aziende poiché CrowdTangle, lo strumento utilizzato per il monitoraggio, non traccia tutti i tipi di account. CrowdTangle monitora oltre 5 milioni di pagine, gruppi e profili verificati su Facebook, inclusi tutti i profili pubblici con più di 25.000 "Mi piace" e gruppi pubblici con più di 95.000 membri. Tuttavia, non supporta il monitoraggio di gruppi Facebook privati, account Instagram privati e profili Facebook personali, a meno che non siano verificati e abbiano il pulsante "Segui" attivato. Questo limita il set di dati alle aziende che rientrano nei criteri di monitoraggio di CrowdTangle, assicurando però che i dati raccolti siano pertinenti.

Inoltre, per i post su Facebook, non è stato possibile misurare né il sentiment né il livello di incertezza. Questo è dovuto al fatto che è stato possibile raccogliere i dati solamente per le 59 aziende tracciate da CrowdTangle. Questo rappresenta uno dei limiti principali dell'analisi effettuata, poiché l'assenza di queste informazioni può influenzare la comprensione del coinvolgimento del pubblico nel processo quotazione in borsa. La possibilità di analizzare il sentiment e il livello di incertezza avrebbe potuto infatti fornire informazioni sulla comunicazione delle aziende e sulla loro capacità di gestire l'avvicinarsi dell'IPO.

3.3.1 Analisi esplorativa delle variabili ricavate da Facebook

In questa sezione vengono analizzate le variabili derivate da Facebook presenti nel dataset. L'obiettivo è comprendere meglio la distribuzione di queste variabili, che possono fornire informazioni utili per le successive analisi. Le tabelle seguenti riassumono le frequenze relative delle variabili categoriali considerate nello studio e le statistiche descrittive delle interazioni totali.

Tabella 3.5: Frequenze relative della variabile In_Fb_List

Categoria	Frequenza Relativa
0	0.4033
1	0.5967

Tabella 3.6: Statistiche descrittive delle interazioni totali su Facebook per 59 aziende

Variabile	Minimo	Q1	Mediana	Media	Q3	Massimo
Interazioni Totali	0	0	0	248.10	0	32094

Tabella 3.7: Frequenze relative della variabile facebook_ipo

Categoria	Frequenza Relativa
0	0.9383
1	0.0617

Le analisi esplorative delle variabili derivate da Facebook forniscono informazioni sulla presenza e l'utilizzo di questa piattaforma social da parte delle aziende oggetto di studio. La Tabella 3.5 mostra che il 59.67% delle aziende possiede una pagina Facebook, mentre il restante 40.33% non ha una presenza ufficiale sulla piattaforma. Questo dato suggerisce che una buona parte delle aziende riconosce l'importanza di Facebook come canale di comunicazione ufficiale. Tuttavia, una

percentuale significativa di aziende non sfrutta questa opportunità.
Le statistiche descrittive delle interazioni totali su Facebook per le 59 aziende, riassunte nella Tabella 3.6, rivelano una notevole variabilità. Il numero di interazioni totali varia da un minimo di 0, per le pagine per cui non è stato possibile scaricare i post, a un massimo di 32.094. La mediana è 0, poiché per più del 50% delle aziende non sono presenti post Facebook nel dataset. La media delle interazioni totali è di 248,10, suggerendo una distribuzione altamente asimmetrica con alcuni valori alti che influenzano la media. Questo potrebbe riflettere differenze nelle strategie di comunicazione, nella visibilità o nell'interesse del pubblico per le diverse aziende.
La Tabella 3.7 presenta le frequenze relative della variabile facebook_ipo. Solo il 6.17% delle aziende ha menzionato l'IPO nei propri post su Facebook nelle due settimane precedenti la quotazione, mentre il 93.83% non ha fatto menzioni dell'IPO. Questo suggerisce che poche aziende sfruttano Facebook per promuovere l'IPO, nonostante la potenziale efficacia della piattaforma per raggiungere un vasto pubblico. La bassa frequenza di menzioni dell'IPO potrebbe indicare una mancanza di fiducia nell'efficacia di questo canale per tale scopo.
In sintesi, l'analisi delle variabili derivate da Facebook rivela che mentre molte aziende riconoscono l'importanza di avere una presenza su questa piattaforma, poche ne sfruttano appieno il potenziale in termini di interazioni e promozione dell'IPO. Questi risultati evidenziano le opportunità per migliorare le strategie di comunicazione digitale delle aziende, soprattutto in relazione agli eventi finanziari chiave come le IPO.

3.3.2 Relazioni tra l'underpricing e le variabili estratte da Facebook

In questa sezione, analizziamo le relazioni tra l'underpricing e alcune variabili estratte da Facebook. L'obiettivo è comprendere se e come il coinvolgimento sui social media possa influenzare il livello di underpricing delle IPO. I box plot riportati di seguito mostrano le distribuzioni dell'underpricing in funzione di queste variabili. La Figura 3.5 mostra il box plot della variabile In_Fb_List in relazione all'underpricing.

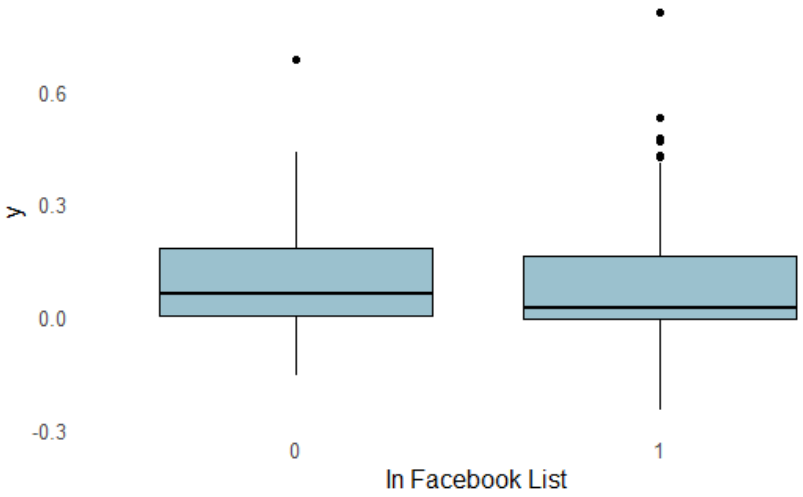

Figura 3.5: Box plot di In Facebook List e Underpricing

La Figura 3.5 mostra che le aziende senza una pagina Facebook tendono ad avere un underpricing leggermente superiore rispetto a quelle che la possiedono. Questo risultato potrebbe essere controintuitivo, poiché ci si potrebbe aspettare che la presenza di una pagina Facebook, e quindi una maggiore visibilità, sia associata a un maggiore interesse del pubblico e a un underpricing più elevato. Tuttavia, ciò potrebbe indicare che altri fattori, oltre alla semplice esistenza della pagina Facebook, influenzano l'underpricing.

È possibile notare che, sebbene le mediane siano abbastanza simili, c'è una maggiore dispersione dei valori e un numero maggiore di outlier per le aziende che dispongono di una pagina Facebook. Questo suggerisce una variabilità più alta nel livello di underpricing per queste aziende.
La Figura 3.6 mostra il box plot della variabile Interazioni_totali e il livello di underpricing. Dato che più del 75% delle interazioni totali sono pari a zero, la variabile è stata suddivisa in due categorie: "Zero Interazioni" e "Interazioni Positive".

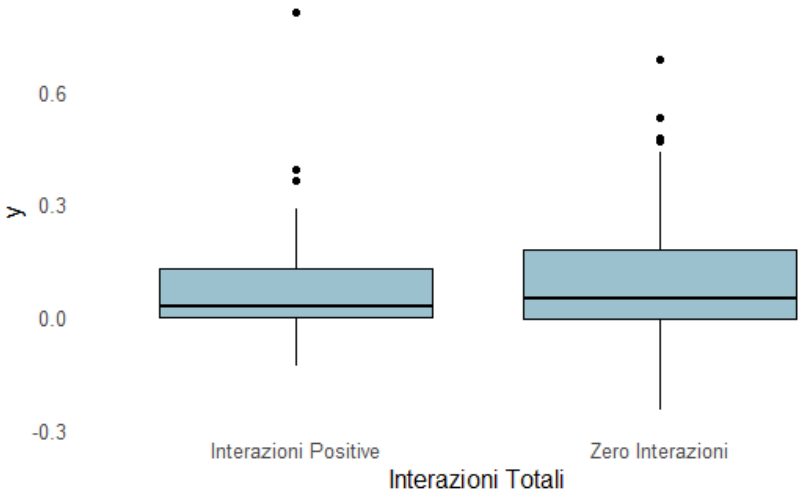

Figura 3.6: Box plot delle Interazioni Totali e Underpricing

La Figura 3.6 mostra che le aziende con interazioni positive tendono ad avere un underpricing leggermente inferiore rispetto alle aziende con zero interazioni. Tuttavia, la differenza tra i due gruppi non appare particolarmente marcata. Questo potrebbe suggerire che, sebbene il coinvolgimento del pubblico sui social media possa avere un impatto sulla risposta, l'effetto potrebbe non essere molto forte.
La Figura 3.7 mostra il box plot della variabile facebook_IPO in relazione all'underpricing. La Figura mostra che le aziende che non hanno menzionato l'IPO nei loro post su Facebook tendono ad avere un underpricing significativamente più elevato rispetto a quelle che hanno menzionato l'IPO. Questo risultato è controintuitivo,

poiché ci si potrebbe aspettare che la promozione dell'IPO sui social media aumenti l'interesse del pubblico e di conseguenza l'underpricing.
Un fattore importante da considerare è che solo il 6.17% delle aziende appartiene alla categoria 1 (hanno menzionato l'IPO su Facebook), il che potrebbe influenzare la distribuzione osservata. La bassa frequenza di aziende che promuovono attivamente l'IPO su Facebook suggerisce che molte aziende potrebbero non sfruttare appieno questa piattaforma per aumentare la visibilità dell'IPO.

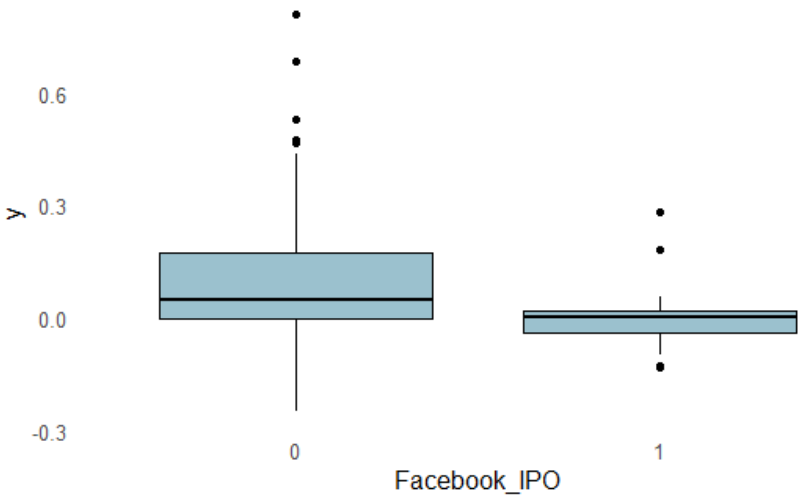

Figura 3.7: Box plot di Facebook IPO e Underpricing

In conclusione, sono stati effettuati i test di Wilcoxon-Mann-Whitney per le variabili Interazioni_totali, In_FB_List e facebook_ipo. L'analisi dei test ha confermato i risultati dell'analisi grafica per la variabile facebook_ipo, il test ha fornito una statistica pari a 2.3587 e p-value = 0.01834. La variabile Interazioni_Totali invece, non ha mostrato differenze significative tra i due gruppi. Per quanto riguarda la variabile In_FB_List, il test ha mostrato l'assenza di differenze tra gruppi ad un livello di significatività dell'5%, ottenendo tuttavia un p-value pari a 0.09297.

3.4 Variabile derivata da Google Trends

La variabile "picco" è una variabile binaria derivata da Google Trends che indica la presenza di un picco di ricerche su Google per ciascuna azienda nei quindici giorni precedenti la data di quotazione in borsa. Il valore è stato impostato a 1 se il picco di ricerche era presente negli ultimi 15 giorni antecedenti la data di quotazione, e a 0 altrimenti. Inoltre, è stato posto a 0 quando la funzione gtrends non è stata in grado di scaricare i dati, una situazione che si verifica quando Google Trends non riesce a creare il grafico a causa di un numero insufficiente di query di ricerca per il termine specifico.
L'obiettivo di questa variabile è identificare un possibile aumento dell'interesse pubblico o dell'attività promozionale immediatamente prima dell'IPO, fornendo un indicatore del livello di attenzione ricevuto dall'azienda. Tuttavia, la mancanza di dati sufficienti per alcune aziende rappresenta un limite dell'analisi, poiché potrebbe influenzare la capacità di rilevare effettivamente un aumento dell'interesse pubblico. La presenza di un picco di ricerche su Google appena prima della quotazione in borsa potrebbe indicare un maggiore coinvolgimento del pubblico.

3.4.1 Analisi esplorativa della variabile ricavata da Google Trends

La Tabella 3.8 mostra le frequenze relative della variabile picco, che indica la presenza di un picco di ricerche su Google per ciascuna azienda nei quindici giorni precedenti la data di quotazione.
La Tabella 3.8 mostra che il 77.78% delle aziende non ha registrato un picco di ricerche su Google nei quindici giorni precedenti la data di quotazione, mentre il 22.22% delle aziende ha registrato un picco di ricerche.

Tabella 3.8: Frequenze relative della variabile Picco

Categoria	Frequenza Relativa
0	0.7778
1	0.2222

Questo risultato suggerisce che la maggior parte delle aziende non sperimenta un aumento significativo dell'interesse del pubblico su Google immediatamente prima della quotazione. Tuttavia, una parte considerevole delle aziende mostra un incremento nelle ricerche, indicando un possibile aumento dell'interesse pubblico.
La bassa frequenza relativa della categoria con picco (valore 1) potrebbe essere dovuta a diversi fattori. Inoltre, la mancanza di dati sufficienti per alcune aziende, dovuta all'incapacità di Google Trends di creare grafici per termini con un numero insufficiente di query di ricerca, rappresenta un limite dell'analisi che potrebbe influenzare la capacità di rilevare effettivamente un aumento dell'interesse pubblico.
In sintesi, la distribuzione delle frequenze relative della variabile picco evidenzia che, mentre la maggior parte delle aziende non registra un aumento significativo delle ricerche su Google prima dell'IPO, una quota significativa mostra un picco di interesse, suggerendo potenziali differenze nell'interesse del pubblico.

3.4.2 Relazione tra l'underpricing e la variabile ricavata da Google Trends

La Figura 3.8 mostra il box plot della variabile picco in relazione all'underpricing.
La Figura 3.8 mostra che le aziende che non hanno registrato un picco di ricerche su Google appena prima della quotazione tendono ad avere un underpricing leggermente superiore rispetto a quelle che hanno registrato un picco di ricerche.
È importante notare che, sebbene le mediane tra i due gruppi siano simili, la dispersione dei valori e il numero di outlier sono maggiori per le aziende senza picco

di ricerche su Google nelle due settimane precedenti l'IPO. Questo indica una maggiore variabilità nell'underpricing per queste aziende, suggerendo che altri fattori potrebbero influenzare l'underpricing oltre all'interesse del pubblico misurato dalle ricerche su Google.
Inoltre, la bassa frequenza relativa delle aziende con picco di ricerche potrebbe influenzare la robustezza delle conclusioni tratte da questa analisi.

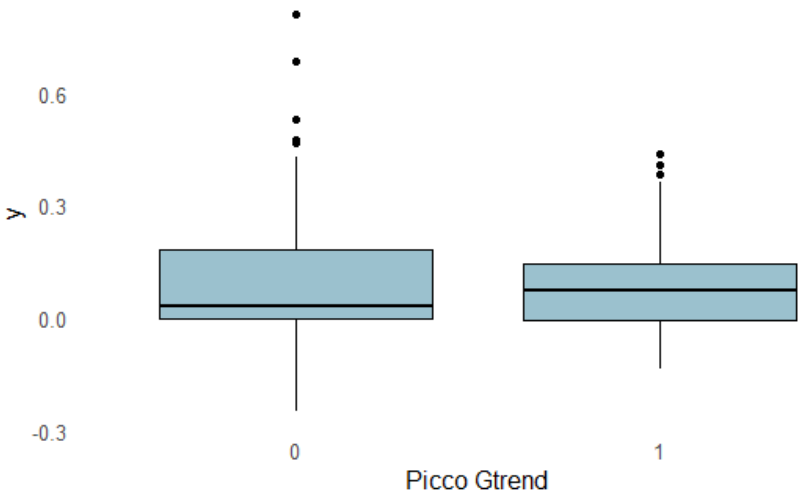

Figura 3.8: Box plot di Picco Gtrend e Underpricing

In conclusione, il box plot suggerisce che le aziende con un picco di ricerche su Google tendono ad avere un underpricing leggermente inferiore rispetto a quelle senza picco. Tuttavia, la maggiore variabilità osservata tra le aziende che non presentano un picco di ricerca nelle due settimane precedenti l'IPO, indica che ulteriori analisi sono necessarie per comprendere i fattori che influenzano l'underpricing e il ruolo dell'interesse pubblico misurato dalle ricerche su Google.
I risultati ottenuti dall'analisi grafica non sono supportati dal test di Wilcoxon-Mann-Whitney.

3.5 Variabili derivate dall'analisi testuale degli articoli

In questa sezione, verrà descritto il processo di estrazione delle variabili dagli articoli finanziari utilizzando tecniche di text mining. Le variabili derivate includono il conteggio totale degli articoli, indici di leggibilità e diversità lessicale, nonché le parole chiave estratte con l'algoritmo RAKE. La sezione successiva invece sarà dedicata all'analisi del sentiment e dell'incertezza negli articoli finanziari.

3.5.1 Pre-processing dei testi e descrizione delle variabili

Il processo di estrazione delle variabili inizia con la lettura dei dati da un file Excel contenente gli articoli finanziari, i quali sono stati aggregati per ciascuna azienda. La prima variabile calcolata è stata il conteggio totale degli articoli per ciascuna azienda, denominata "Articoli_Totali". Questo conteggio è stato ottenuto raggruppando gli articoli per azienda e calcolando il numero totale di articoli presenti per ciascuna.
Successivamente, i testi sono stati tokenizzati, ovvero suddivisi in unità più piccole chiamate token, che possono essere parole, frasi o altri elementi significativi. Utilizzando il pacchetto `tidytext`, il testo è stato suddiviso in singole parole, rimuovendo numeri, simboli, punteggiatura e URL. Le stop words italiane, ovvero le parole vuote che non aggiungono valore significativo all'analisi, sono state rimosse, insieme a termini specifici che non erano utili per l'analisi, come codici e identificatori.
Gli indici di leggibilità sono stati calcolati utilizzando la funzione textstat_readability del pacchetto `quanteda`. In particolare, sono stati calcolati due indici:

- Lunghezza media delle frasi ("meanSentenceLength"): questo indice è calco-

lato come il rapporto tra il numero totale di frasi n_{st} e il numero di parole n_w. La formula utilizzata è:

$$\text{ASL} = \frac{n_{st}}{n_w} \tag{3.2}$$

- Media delle sillabe per parola ("meanWordSyllables"): questo indice è calcolato come il rapporto tra il numero di parole n_w e il numero di sillabe n_{sy}. La formula utilizzata è:

$$\text{AWL} = \frac{n_w}{n_{sy}} \tag{3.3}$$

La diversità lessicale è stata calcolata utilizzando la funzione textstat_lexdiv del pacchetto `quanteda`. Per questa analisi, è stato utilizzato il parametro "R" di Guiraud's Root TTR (Type-Token Ratio), che fornisce una misura della diversità lessicale nei testi. L'indice "R" è definito come il rapporto tra il numero totale di token (N) e la radice quadrata del numero di types (V). La formula utilizzata è:

$$R = \frac{N}{\sqrt{V}} \tag{3.4}$$

L'algoritmo RAKE (Rapid Automatic Keyword Extraction) è stato applicato ai testi puliti per estrarre le parole chiave più rilevanti. RAKE è un metodo di estrazione automatica delle parole chiave che identifica termini e frasi che appaiono frequentemente nel testo e sono significativi nel contesto in esame.
Come presentato da Rose et al. (2010), il processo di estrazione delle parole chiave inizia con la suddivisione del testo in parole candidate. In seguito, dopo aver ottenuto la lista delle parole chiave candidate e completato il grafo delle co-occorrenze, viene calcolato un punteggio per ciascuna parola. Il punteggio di ciascuna parola viene calcolato per ciascun testo presente nel corpus. Questo punteggio, viene ottenuto come il rapporto tra il grado delle parole deg(w) e la frequenza delle parole freq(w). Durante questo processo, le parole vuote italiane non vengono rimosse,

sono infatti fondamentali per estrarre le parole chiave. Le parole vuote delimitano i confini delle frasi e aiutano a preservare i pattern di co-occorrenza per l'estrazione delle parole chiave.
In questo studio la ricerca si è limitata a considerare esclusivamente gli unigrammi, cioè singole parole. Inoltre, le parole estratte sono state filtrate considerando solamente quelle che sono apparse con una frequenza minima pari a 10. Questa soglia di frequenza è stata impostata in modo tale che solo le parole chiave più significative e frequenti fossero estratte dal corpus.
In conclusione, al termine di questo processo, sono state identificate le 50 parole chiave con indice RAKE più elevato, ponendole quindi come nuove variabili all'interno del dataset.

3.5.2 Analisi esplorativa univariata delle variabili estratte

In questa sezione vengono analizzate le statistiche descrittive delle variabili quantitative estratte dagli articoli finanziari. La Tabella 3.9 riassume le principali caratteristiche di queste variabili, inclusi i valori di minimo, primo quartile (Q1), mediana, media, terzo quartile (Q3) e massimo.

Tabella 3.9: Statistiche descrittive delle variabili estratte dagli articoli finanziari

Variabile	Minimo	Q1	Mediana	Media	Q3	Massimo
Articoli_Totali	0.000	1.000	3.000	4.231	5.000	30.000
R	0.000	8.881	11.635	11.486	14.433	27.268
meanWordSyllables	0.000	2.158	2.228	1.937	2.283	2.425
meanSentenceLength	0.000	21.429	26.042	24.430	30.315	87.500

La Tabella 3.9 mostra che la distribuzione delle variabili quantitative presenta una variabilità significativa tra le aziende. La variabile "Articoli_Totali" indica il numero totale di articoli pubblicati per ciascuna azienda, con una media di circa

4,23 articoli e un massimo di 30 articoli. L'indice "R" misura la diversità lessicale nei testi, con una mediana di circa 11,64 e un massimo di 27,27. La variabile "meanWordSyllables" riflette la complessità delle parole nei testi, con una media di circa 1,94 sillabe per parola. Infine, la variabile "meanSentenceLength" fornisce indicazioni sulla leggibilità del contenuto degli articoli, con una media di circa 24,43 parole per frase e un massimo di 87,50 parole.
La Figura 3.9 mostra le 20 parole chiave con il punteggio RAKE più elevato.

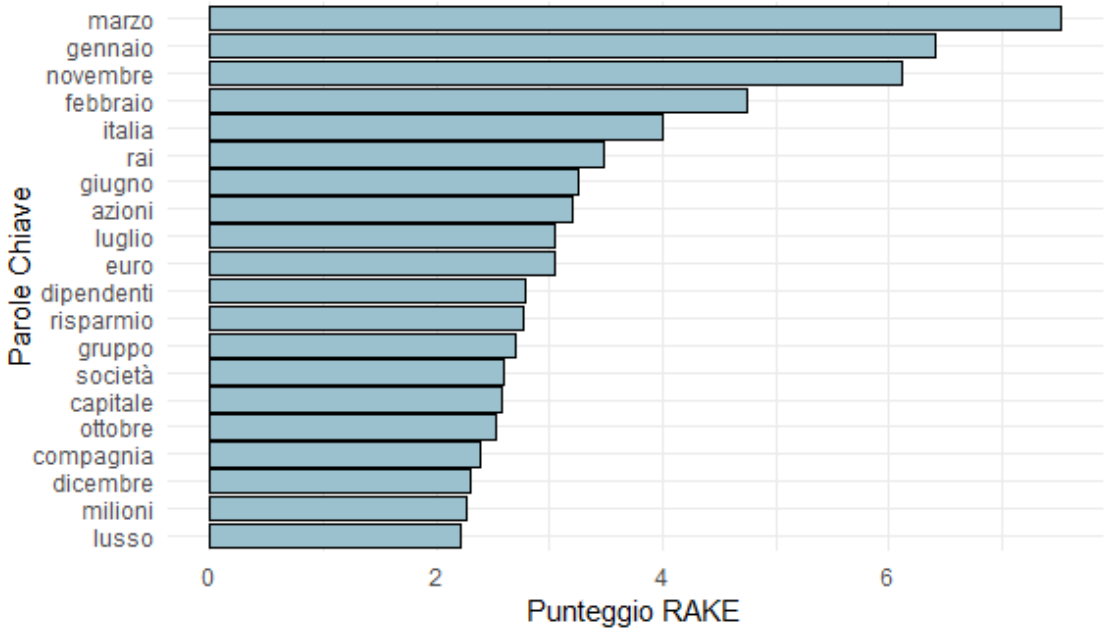

Figura 3.9: Frequenza delle Parole Chiave Selezionate

Queste parole chiave sono quelle che appaiono frequentemente nei testi e sono significative nel contesto, evidenziando i temi più rilevanti trattati negli articoli finanziari. Parole come "marzo", "gennaio" e "novembre" suggeriscono un'alta concentrazione di notizie legate a specifici periodi dell'anno. Altri termini come "italia", "rai", "azioni" e "euro" indicano la presenza di argomenti aziendali rilevanti.
La Figura 3.10 mostra la frequenza delle parole chiave selezionate. Questo gra-

fico evidenzia quali parole chiave sono più comuni nel corpus, fornendo ulteriori informazioni sulla rilevanza e la diffusione di specifici termini nei testi.

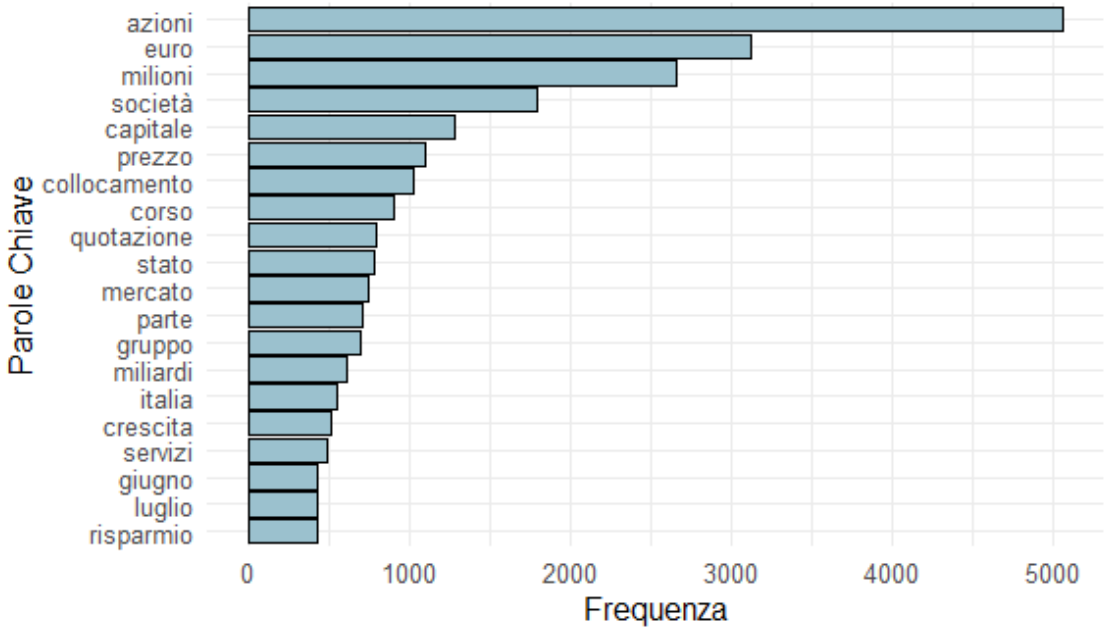

Figura 3.10: Frequenza delle Parole Chiave Selezionate

La Figura 3.10 visualizza la frequenza delle principali parole chiave selezionate in base al punteggio RAKE. È evidente come termini finanziari e aziendali siano più presenti. La parola "azioni" che emerge come la parola chiave più frequente, è seguita da "euro" e "milioni".

Altri termini di rilievo includono "società", "capitale" e "prezzo", riflettendo una focalizzazione su questioni aziendali.

È importante notare che per 32 aziende non sono stati estratti articoli. Di conseguenza, i valori nel dataset per queste aziende sono pari a zero, il che ha un'influenza significativa sulle distribuzioni delle variabili. Questo aspetto deve essere tenuto in considerazione nell'interpretazione dei risultati, poiché la presenza di un numero considerevole di aziende senza articoli può distorcere le statistiche descrittive e for-

nire una rappresentazione non completamente accurata della variabilità nei testi analizzati.

3.5.3 Relazioni tra l'Underpricing e le variabili estratte dagli articoli finanziari

In questa sezione, viene analizzata la relazione tra l'underpricing e le variabili quantitative estratte dagli articoli finanziari. La Figura 3.11 mostra gli scatter plot delle variabili "Articoli_Totali", "R", "meanWordSyllables" e "meanSentenceLength" in relazione all'underpricing.
La Figura 3.11 evidenzia le relazioni tra l'underpricing e le variabili quantitative estratte dagli articoli finanziari. Dall'analisi dei grafici, si può osservare che:

- La variabile "Articoli_Totali" mostra una distribuzione molto concentrata verso lo zero, con una leggera tendenza positiva rispetto all'underpricing. Questo suggerisce che un maggior numero di articoli potrebbe essere associato a un underpricing leggermente più elevato, anche se la relazione non è forte.

- L'indice "R" non mostra una chiara relazione con l'underpricing, indicando che la diversità lessicale nei testi potrebbe non influenzare significativamente l'underpricing.

- La variabile "meanWordSyllables" non presenta una relazione evidente con l'underpricing, suggerendo che la complessità delle parole nei testi non ha un impatto diretto sull'underpricing.

- La variabile "meanSentenceLength" mostra una leggera relazione negativa con l'underpricing, indicando che articoli con frasi più lunghe potrebbero essere associati a un underpricing leggermente inferiore.

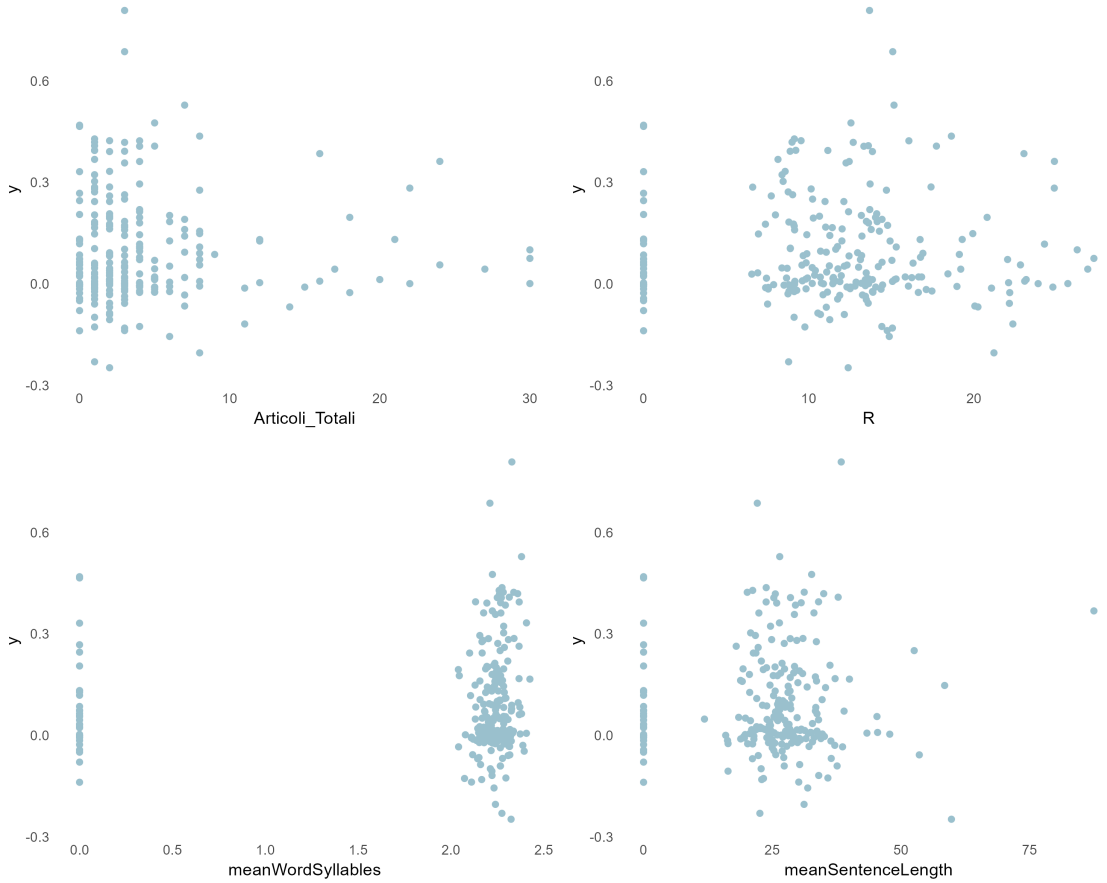

Figura 3.11: Scatter plot delle variabili quantitative estratte dagli articoli finanziari in relazione all'underpricing

In conclusione, sono stati effettuati i test di correlazione di Spearman, che non hanno mostrato significatività per nessuna delle variabili quantitative estratte dagli articoli finanziari.

La Figura 3.12 mostra il coefficiente di correlazione di Spearman tra le 20 correlazioni più alte in valore assoluto tra la variabile di risposta y (underpricing) e le 50 parole estratte tramite indice RAKE.

Dall'analisi del grafico emergono alcune osservazioni. In primo luogo, termini come "fiducia", "ottobre" e "azioni" mostrano una correlazione positiva con l'underpricing. Questo suggerisce che la presenza di questi termini negli articoli potrebbe

essere associata a un underpricing più elevato. In particolare, il termine "fiducia" ha la correlazione positiva più forte.

In secondo luogo, alcune parole chiave mostrano una correlazione negativa con l'underpricing. Parole come "tramite", "aprile" e "marzo" indicano che la loro presenza negli articoli potrebbe essere associata a un underpricing inferiore.

È importante notare che tutti i livelli di correlazione presentati sono piuttosto bassi. Questo implica che, sebbene alcune parole chiave mostrino correlazioni leggermente positive o negative con l'underpricing, la forza di queste relazioni è debole. La bassa correlazione indica che le parole chiave, da sole, non sono forti predittori dell'underpricing.

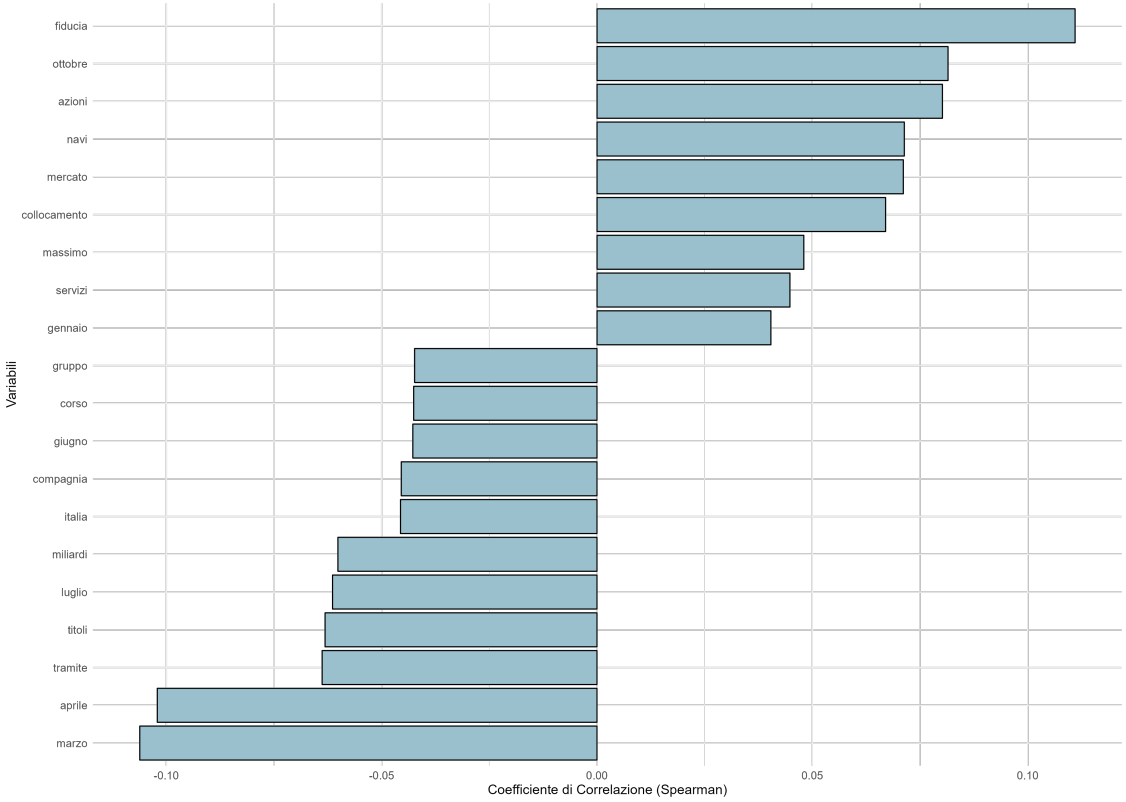

Figura 3.12: Correlazione tra Parole Chiave Selezionate e Underpricing

In conclusione, sono stati effettuati i test di correlazione di Spearman, che non hanno mostrato significatività per nessuna delle variabili estratte tramite indice RAKE.
È importante far notare nuovamente che per 32 aziende non sono stati estratti articoli, ottenendo valori pari a zero per queste osservazioni. Questo aspetto, già evidenziato nell'analisi esplorativa univariata, influenza significativamente la distribuzione delle variabili. Tuttavia, è rilevante anche considerare che la mancanza di articoli per queste aziende potrebbe riflettere una minore copertura mediatica o interesse da parte della stampa finanziaria, il che potrebbe a sua volta influenzare l'underpricing delle rispettive IPO.

3.6 Sentiment Analysis

La sentiment analysis rappresenta una disciplina fondamentale per l'estrazione e l'identificazione delle emozioni espresse nei testi, siano esse positive, negative o neutre. Questo campo di studio è di particolare interesse per decisori quali politici, amministratori, manager aziendali e ricercatori nel settore delle scienze sociali, poiché offre strumenti per comprendere l'opinione pubblica su vari temi rilevanti. Con l'avvento di Internet e dei social media, la quantità di dati disponibili per l'analisi del sentiment è aumentata in maniera esponenziale, fornendo così fonti inesauribili di informazioni preziose (Ceron, et al. (2014)).
L'analisi testuale si basa su alcuni principi fondamentali: ogni modello quantitativo linguistico è sbagliato, ma può risultare utile in contesti specifici; sebbene i metodi automatici migliorino l'efficienza dell'analisi, non possono sostituire completamente la comprensione umana del linguaggio; infine, non esiste una tecnica universale per l'analisi testuale, poiché ogni metodo ha applicazioni specifiche e vincoli propri.
Come presentato da Ceron, et al. (2014), si possono citare i seguenti metodi di clas-

sificazione nell'analisi del sentiment, suddivisi in metodi supervisionati e non supervisionati: il clustering, una tecnica non supervisionata che raggruppa le osservazioni in sottogruppi omogenei basati su misure di dissimilarità; e il machine learning, che comprende metodi supervisionati come Support Vector Machines, Random Forests e Neural Networks, i quali classificano i dati basandosi su un set di addestramento precodificato.

Nell'ambito della sentiment analysis, un metodo di classificazione frequentemente utilizzato è basato su dizionari ontologici. Questi dizionari consistono in raccolte di termini categorizzati secondo macro-aree semantiche. Questi metodi si rivelano particolarmente efficaci quando i vincoli sono ben definiti e il rumore è minimo. Tuttavia, in contesti come i social media, dove il linguaggio è variegato e spesso rumoroso, possono presentare limitazioni in termini di precisione.

Un'importante innovazione nell'analisi del sentiment è rappresentata dall'algoritmo iSA (integrated Sentiment Analysis). L'Integrated Sentiment Analysis (iSA) combina dati da diverse fonti (social media, recensioni, notizie) e applica vari modelli di sentiment analysis (supervisionati, non supervisionati, basati su lexicon) a ciascuna fonte. I risultati individuali vengono poi integrati per ottenere una valutazione complessiva più accurata e rappresentativa del sentiment. Questo approccio migliora la robustezza e l'affidabilità dell'analisi del sentiment rispetto ai metodi tradizionali.

3.6.1 I dizionari ontologici di Loughran e McDonald

Un importante contributo nell'ambito dell'analisi del sentiment nei testi finanziari tramite dizionari ontologici è stato apportato da Loughran e McDonald. Nel loro studio, gli autori hanno dimostrato che i dizionari di parole negative sviluppati per altre discipline spesso, classificano erroneamente come negative parole comuni in contesti finanziari. Per ovviare a questo problema, hanno sviluppato dei dizionari ontologici specifici per i testi finanziari, noti come Fin-Neg (Negative), Fin-Pos

(Positive), Fin-Unc (Uncertainty) e Fin-Lit (Litigious).
In questo lavoro, sono stati utilizzati i dizionari proposti da Loughran e McDonald per calcolare il sentiment (utilizzando Fin-Neg e Fin-Pos) e l'incertezza (utilizzando Fin-Unc) nei testi analizzati. Nello specifico, i dizionari impiegati sono stati tradotti in italiano dal Professor Andrea Sciandra dell'Università di Padova e dal Professor Riccardo Ferretti dell'Università di Modena Reggio Emilia (Ferretti e Sciandra (2022)). Queste traduzioni hanno permesso di applicare le categorie semantiche definite da Loughran e McDonald su testi finanziari in lingua italiana, mantenendo la stessa accuratezza e rilevanza semantica dei dizionari originali.

3.6.2 Analisi esplorativa del sentiment e dell'incertezza

La sentiment analysis è stata condotta utilizzando i dizionari ontologici di Loughran e McDonald tradotti in italiano. Per calcolare i valori di sentiment, sono stati utilizzati i dizionari delle parole positive e negative, che contribuiscono a determinare il sentiment complessivo dei testi finanziari analizzati. I risultati sono stati memorizzati nella variabile Net_Sentiment. Per l'analisi dell'incertezza, è stato utilizzato il dizionario specifico per le parole incerte. I risultati ottenuti sono stati memorizzati nella variabile Net_Uncertain.
I grafici delle parole positive e negative in Figura 3.13 mostrano la frequenza delle parole più ricorrenti classificate come positive e negative nei testi analizzati. Questi grafici sono fondamentali per comprendere il contributo di queste parole al sentiment complessivo.

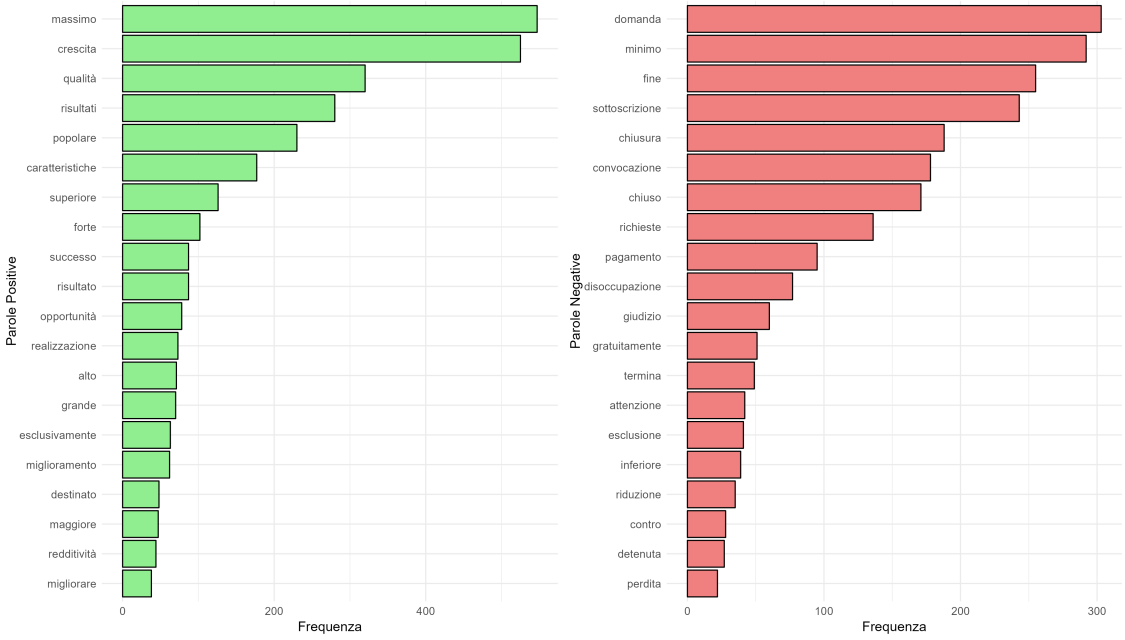

Figura 3.13: Grafico delle Parole Positive e Negative

Il grafico delle parole positive mostra le parole più frequenti che esprimono sentimenti positivi nei testi finanziari. Parole come "massimo", "crescita", "qualità" e "risultati" sono tra le più ricorrenti.

Il grafico delle parole negative evidenzia le parole più frequenti che esprimono sentimenti negativi. Termini come "domanda", "minimo", "fine" e "sottoscrizione" sono comuni.

Il grafico delle parole incerte mostra la frequenza delle parole più ricorrenti classificate come incerte nei testi analizzati. Questo grafico aiuta a identificare il linguaggio che esprime incertezza e contribuisce al calcolo dell'incertezza complessiva nei testi finanziari, i cui risultati sono memorizzati nella variabile Net_Uncertain.

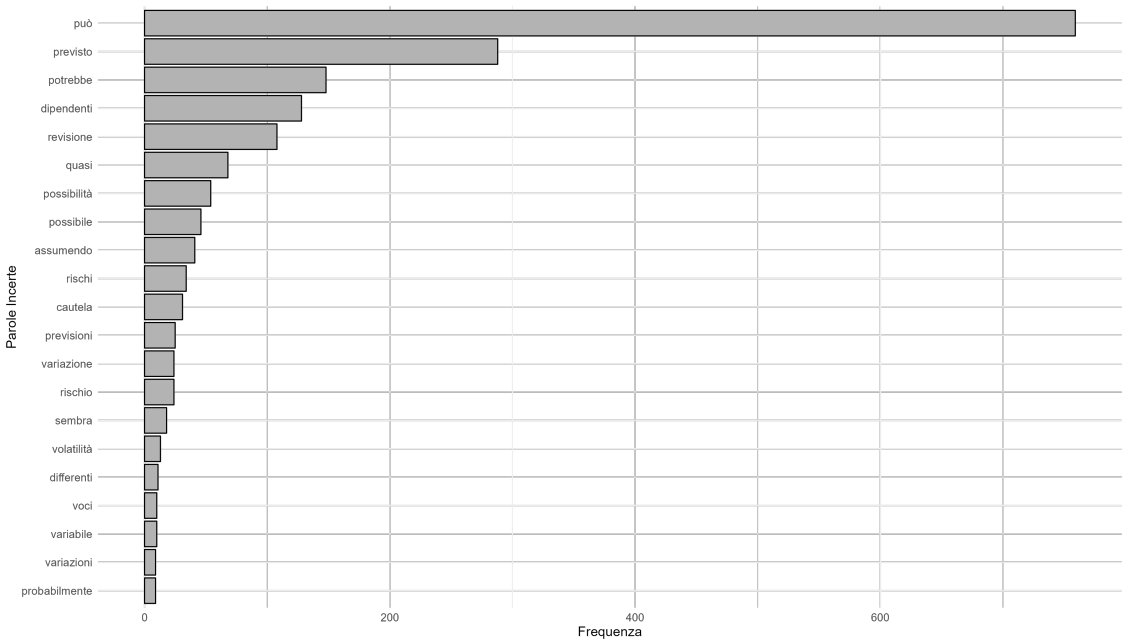

Figura 3.14: Grafico delle Parole Incerte

Il grafico delle parole incerte include termini come "può", "previsto", "potrebbe", "dipendenti" e "revisione". Queste parole indicano incertezza, possibilità e variabilità nelle previsioni e nelle dichiarazioni. L'uso frequente di tali termini suggerisce che nei testi analizzati ci sono molte dichiarazioni che lasciano spazio a incertezze. L'analisi delle parole più frequenti nei testi finanziari, suddivise per sentiment positivo, negativo e incerto, fornisce una panoramica del linguaggio utilizzato per trasmettere emozioni e comunicare incertezze.

3.6.3 Relazioni tra underpricing, sentiment e incertezza

La Figura 3.15 mostra gli scatter plot tra il livello di underpricing (y) e le variabili relative al sentiment e all'incertezza.

Figura 3.15: Scatter plot tra underpricing (y) e le variabili Sentiment e Incertezza

Dall'analisi dei grafici, emergono i seguenti aspetti rilevanti:

- La distribuzione dei punti nello scatter plot tra sentiment e underpricing appare abbastanza diffusa senza un chiaro trend. Questo suggerisce che non c'è una relazione forte e diretta tra il sentiment e l'underpricing. Sono presenti alcuni outliers, sia positivi che negativi, che si distaccano dalla massa dei dati. Tuttavia, questi outliers non sembrano seguire un pattern specifico che colleghi in modo evidente il sentiment all'underpricing.

- La distribuzione dei punti nello scatter plot tra incertezza e underpricing è anch'essa diffusa senza un chiaro trend evidente. Alcuni punti si distaccano dal resto, indicando un livello più alto di incertezza o di underpricing. Tuttavia, come per il sentiment, non mostrano un pattern specifico che colleghi chiaramente l'incertezza all'underpricing.

I risultati ottenuti dall'analisi grafica sono supportati dal test di correlazione di Spearman, che mostra assenza di correlazione tra sentiment, uncertain e underpricing.

Capitolo 4

Analisi e implementazione dei modelli di Machine Learning

In questo capitolo, verranno presentati i tre modelli di machine learning utilizzati per analizzare l'underpricing delle IPO: Lasso, Random Forest e Support Vector Regression. L'obiettivo è identificare i fattori chiave che influenzano le variazioni nei rendimenti delle IPO e sviluppare una capacità predittiva accurata riguardo a questi rendimenti, migliorando così la comprensione delle dinamiche finanziarie sottostanti.

Il capitolo è organizzato come segue: nella Sezione 4.1 verrà esaminata la letteratura relativa ai modelli proposti, nella Sezione 4.2 saranno descritte le operazioni preliminari effettuate sui dati, e nella Sezione 4.3 verrà presentata l'analisi empirica dei modelli Lasso, Random Forest e Support Vector Regression. Infine, nella Sezione 4.4 verranno confrontati i modelli in termini di RMSE e MAE, fornendo una valutazione delle loro performance predittive.

4.1 Revisione della letteratura sui modelli

In questa sezione, verranno brevemente presentati da un punto di vista teorico i tre modelli di machine learning proposti per l'analisi dell'underpricing delle IPO: Lasso, Random Forest e Support Vector Regression.

4.1.1 Lasso

La Regressione Lasso (Least Absolute Shrinkage and Selection Operator) è un metodo utilizzato per effettuare shrinkage e selezionare automaticamente le variabili significative in un modello di regressione lineare imponendo una penalizzazione L1. La regressione Lasso è un metodo molto simile a quello dei minimi quadrati ordinari, e come illustrato da Tibshirani (1996) risolve il seguente problema di minimo:

$$\min_{\beta \in \mathbb{R}^p} (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|_1$$

Il parametro di ottimizzazione per questo modello è λ. Questo parametro controlla la quantità di regolarizzazione applicata ai coefficienti β_j. Un valore maggiore di λ implica una penalizzazione più forte, che può ridurre alcuni coefficienti β_j a zero, effettuando una selezione delle variabili e mantenendo solo quelle più significative.
Tibshirani (1996) identifica quindi due ragioni principali per cui le stime Lasso potrebbero essere più soddisfacenti rispetto alle stime OLS. La prima riguarda l'accuratezza delle previsioni: le stime OLS spesso presentano un bassa distorsione ma una grande varianza. L'accuratezza delle previsioni può essere migliorata riducendo alcuni coefficienti β_j a zero o diminuendone l'entità. In questo modo, si accetta un aumento della distorsione nelle stime dei coefficienti per ridurre la loro variabilità, migliorando così l'accuratezza complessiva delle previsioni. La seconda ragione ri-

guarda l'interpretazione. Con un gran numero di predittori, è spesso desiderabile identificare un sottoinsieme più ristretto che mostri gli effetti più rilevanti.

4.1.2 Random Forest

Il Random Forest è un algoritmo di apprendimento automatico che combina un insieme di alberi decisionali per migliorare la precisione delle previsioni e ridurre il rischio di overfitting. Questo metodo rientra nella categoria degli algoritmi di "ensemble learning", dove il risultato finale è ottenuto aggregando le previsioni di molti modelli deboli per produrre un modello robusto.
Inizialmente vengono adattati molti alberi di regressione a ciascun campione bootstrap. Questi campioni sono costruiti casualmente con reinserimento dal dataset d'addestramento. La costruzione di ogni albero è molto veloce perché si utilizzano poche variabili ad ogni nodo. Per ciascun albero, la suddivisione che minimizza un determinato criterio di splitting viene scelto come punto di suddivisione del nodo. Il valore previsto per un osservazione viene in seguito calcolato facendo la media dei valori di tutti gli alberi. Per una visione più completa dell'algoritmo di stima si faccia riferimento all'articolo di Biau e Scornet (2016).
Tuttavia, come riportato nell'articolo di Boulesteix et al. (2019), gli iperparametri e le strategie di ottimizzazione in un modello Random Forest possono essere diverse:

- Mtry: numero di variabili candidate in ciascuna suddivisione.
- Sample Size: numero di osservazioni per ogni albero.
- Min_Node_Size: numero minimo di osservazioni per ciascuna foglia.
- Number_Of_Trees: numero di alberi nella foresta.
- Splitting_Rule: criterio di splitting per ciascuna suddivisione.

Il Random Forest presenta numerosi vantaggi, resi ancora più evidenti dall'ottimizzazione degli iperparametri:

- Robustezza ai dati rumorosi: Grazie alla media delle previsioni di molti alberi, il modello è meno sensibile a variazioni nei dati e tende a essere più stabile.

- Capacità di gestire grandi set di dati: Può lavorare efficacemente con dataset di grandi dimensioni.

- Importanza delle variabili: Utilizzando l'errore OOB, il Random Forest fornisce inforomazini su quali variabili hanno il maggiore impatto sulle previsioni.

4.1.3 Support Vector Regression

Il Support Vector Machine è un algoritmo di apprendimento automatico inizialmente sviluppato per problemi di classificazione e successivamente esteso per problemi di regressione, noto come Support Vector Regression. Gli SVR cercano di trovare la funzione che approssima al meglio i dati di addestramento mantenendo l'errore di previsione entro un certo margine.
Il problema di ottimizzazione del Support Vector Regression (SVR), come presentato da Basak et al. (2007), è formulato come segue:

$$\min_{w,b,\xi,\xi^*} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{soggetto ai vincoli} \quad \begin{cases} y_i - \langle w, x_i\rangle - b \leq \epsilon + \xi_i, \\ \langle w, x_i\rangle + b - y_i \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \end{cases}$$

dove:

- Il primo termine, $\frac{1}{2}\|w\|^2$, mira a minimizzare la norma del vettore dei pesi w, cercando di trovare una funzione il più piatta possibile,
- ϵ è il margine di tolleranza,
- ξ_i e ξ_i^* sono variabili di slack,
- C è il fattore di penalizzazione.

Anche per questo modello, come presentato dal problema di ottimizzazione soggetto ai vincoli e descritto da Ito e Nakano (2003), esistono diversi iperparametri e strategie di ottimizzazione:

- ϵ: determina la larghezza della "fascia di tolleranza" attorno alla funzione di regressione.
- C: fattore di penalizzazione per ciascuna deviazione al di fuori della "fascia di tolleranza".
- Kernel Function: la selezione della funzione kernel e del parametro ad esso associato hanno un impatto importante sui modelli finali, permettendo infatti di trasformare i dati originali in uno spazio ad alta dimensionalità, dove diventa possibile catturare relazioni non lineari.

4.2 Operazioni preliminari

Sono state eseguite alcune operazioni preliminari prima di effettuare la fase di modellazione statistica. Per la gestione di questa fase e la successiva di modellazzione è stata seguita la pipeline proposta dal pacchetto `caret` in R:

- La variabile Sponsor è stata ricodificata in 14 variabili binarie, una per ciascuno sponsor tranne Intermonte SIM, al fine di evitare problemi di multicollinearità.

- Il dataset è stato suddiviso in una porzione di addestramento (80%) e una porzione di test (20%). Il dataset di addestramento dispone di 195 righe e 82 colonne, quello di test di 48 righe e 82 colonne. Questa suddivisione consente di addestrare il modello su una porzione dei dati e valutarne le prestazioni su dati non utilizzati in fase di addestramento.

- Le variabili quantitative sono state centrate attorno alla media e scalate secondo la deviazione standard. Questo è importante per evitare che le variabili con scale diverse influenzino in modo sproporzionato il modello.

- Tramite la funzione fitControl è stato definito il processo di addestramento del modello. Si è specificato l'utilizzo della convalida incrociata a 5 fold come metodo di valutazione dei modelli. Si è inoltre specificato il parametro "tolerance", che seleziona il modello con il valore di performance peggiore entro l'1% del migliore.

4.3 Implementazione dei modelli

In questa sezione verranno presentati i risultati ottenuti nella fase di modellazione utilizzando i tre modelli proposti: Lasso, Random Forest e Support Vector Regression. Verranno discussi i parametri utilizzati, le prestazioni dei modelli e le interpretazioni dei risultati.

4.3.1 Lasso

La griglia dei parametri è stata definita specificando un valore fisso di alpha pari a 1, che corrisponde al metodo di regolarizzazione Lasso. Il parametro lambda è stato lasciato libero di variare tra 0 e 0.1, creando una serie di valori su cui verrà effettuata la ricerca del miglior modello.

Per l'addestramento del modello, è stato utilizzato il metodo glmnet. Questa configurazione permette di identificare il valore ottimale di lambda mantenendo alpha fisso a 1, ottenendo così il modello Lasso ottimale.
Il grafico in Figura 4.1 mostra l'andamento dell'indice RMSE rispetto ai valori di lambda. Il valore ottimale ottenuto per lambda è pari a 0.014. Dal grafico in esame possiamo notare come il valore selezionato non corrisponda esattamente al valore che minimizza l'RMSE. Questo avviene perché, come presentato precedentemente, la scelta del parametro è impostata sulla base di un criterio di tolleranza.

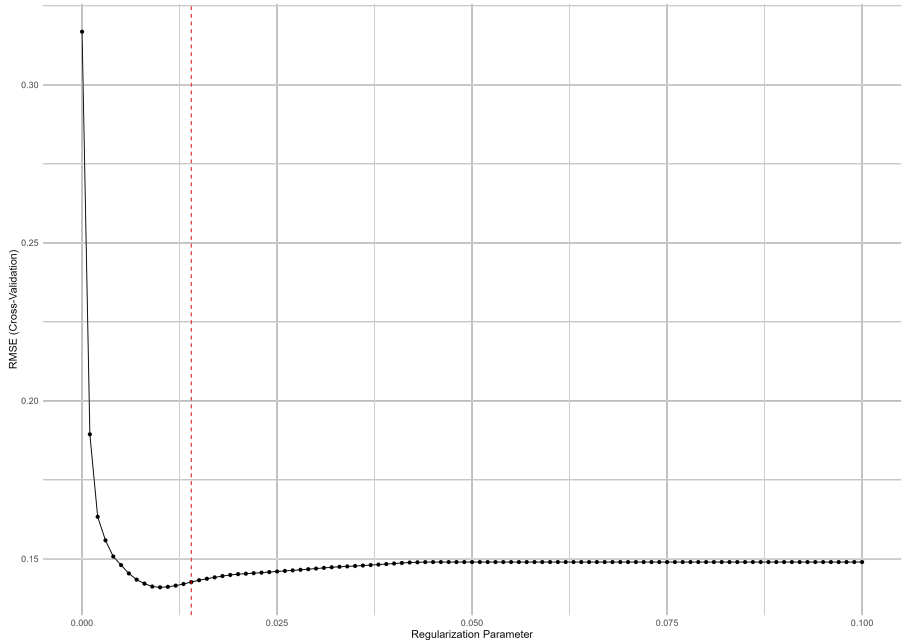

Figura 4.1: Andamento del RMSE in funzione di lambda.

Nel modello stimato, il numero di parametri diversi da zero selezionati con il valore ottimale di lambda è pari a 9, esclusa l'intercetta. Questo numero indica che, su un totale di 82 variabili, solo 9 variabili sono state considerate significative dal modello.

I parametri non nulli possono essere suddivisi in positivi e negativi, come riportato nella Tabella 4.1.

Tabella 4.1: Coefficiente dei parametri non nulli nel modello Lasso

Parametro	**Coefficiente**
Parametri Positivi	
Technology	0.7778
Alantra Capital Markets	0.0113
Banca Profilo	0.0626
BPER Banca	0.0727
lusso	0.0134
Parametri Negativi	
In_FB_List	-0.0079
facebook_ipo	-0.0121
Banca Popolare di Vicenza	-0.0283
miliardi	-0.0018

Nella Tabella 4.1, i parametri positivi e negativi sono stati distinti per facilitare l'interpretazione del modello. I parametri positivi sono quelli con un effetto diretto positivo sulla variabile di risposta e si rifanno a:

- Variabile economica: Technology,
- Sponsor: Alantra Capital Markets, Banca Profilo e BPER Banca,
- Parola: Lusso

I parametri negativi hanno un effetto diretto negativo e si rifanno a:

- Variabili legate a Facebook: In_FB_List e facebook_ipo
- Sponsor: Banca Popolare di Vicenza
- Parola: miliardi

Questi risultati confermano le indicazioni ottenute dalle analisi esplorative sull'impatto positivo legato al settore Technology e agli sponsor, e l'impatto negativo delle variabili legate a Facebook. In particolare, si può osservare che la variabile che più influenza il rendimento al primo giorno di mercato è l'appartenenza al settore Technology, con un coefficiente pari a 0.7778, indicando che appartenere a questo settore aumenta l'underpricing di tale valore.

4.3.2 Random Forest

La griglia dei parametri per il modello Random Forest è stata definita con i seguenti parametri di tuning:

- mtry, rappresenta il numero di variabili da considerare ad ogni split e varia da 1 a 50.
- min.node.size, indica la dimensione minima dei nodi terminali ed è stato impostato a 3, 5 e 7.
- splitrule, determina la regola di split ed è stata impostata a "variance" e "extratrees".

Il criterio di split "variance" crea split che riducono la varianza all'interno di ciascun nodo, aumentando cosi la purezza dei nodi stessi. Il criterio "extratrees" (Extreme Randomized Trees) introduce un ulteriore livello di causalità scegliendo casualmente i punti di divisione. Questa regola può portare a una maggiore diversità tra gli alberi e potenzialmente migliorare la generalizzazione del modello.
Per l'addestramento del modello, è stato utilizzato il pacchetto `ranger` con 300 alberi e l'importanza delle variabili calcolata tramite la varianza delle risposte.
Il grafico in Figura 4.2 mostra l'andamento dell'indice RMSE rispetto al numero di variabili considerate per ogni split e alla dimensione minima dei nodi terminali del

modello Random Forest. Le linee colorate rappresentano i diversi valori del numero minimo di osservazioni per i nodi terminali, mentre l'asse delle ascisse mostra i valori del numero di variabili considerate per ogni split. Il grafico è diviso in due parti: una per la regola di split "variance" e l'altra per "extratrees".

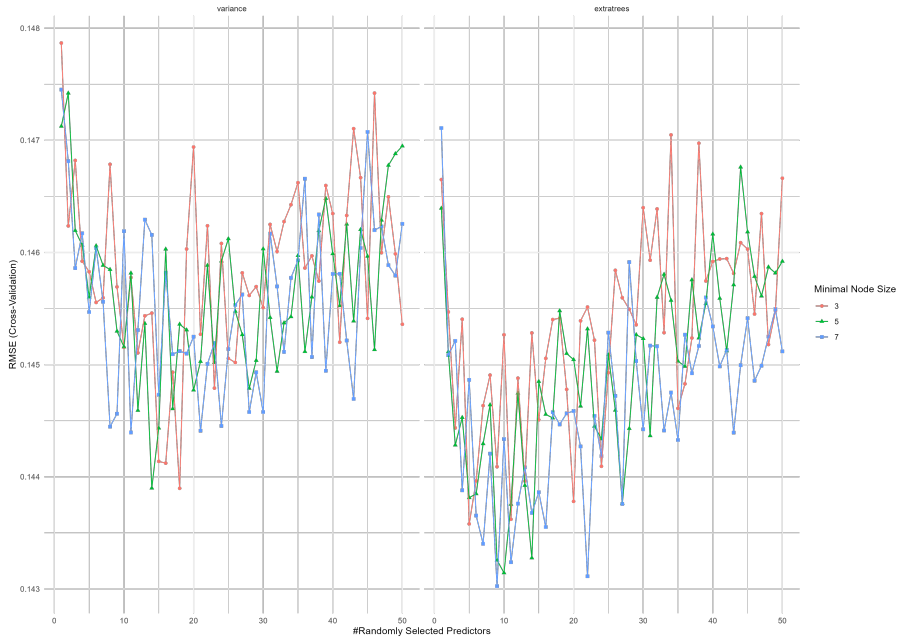

Figura 4.2: Andamento del RMSE in funzione del numero di mtry e min.obs.size.

Dall'output del modello, è stato ottenuto come miglior modello secondo il criterio di tolleranza quello che presenta:

- Numero di variabili candidate per ciascuna suddivisione pari a 2.
- Numero minimo di osservazioni per ciascuna foglia pari a 5.
- Criterio di splitting "extratrees".

Il grafico in Figura 4.3 mostra l'importanza delle 20 variabili più significative nel modello Random Forest. Queste variabili sono considerate le più rilevanti dal

modello per prevedere l'underpricing delle IPO. Coerentemente con quanto ottenuto nell'analisi esplorativa e dal modello Lasso, notiamo che le 3 variabili con un importanza maggiore sono Technology, In_FB_List e BPER Banca. Notiamo inoltre un'importanza associata alle variabili ottenute dall'analisi testuale degli articoli finanziari. Risultano infatti importanti sia le variabili associate alle parole sia indicatori quali meanSentenceLength, Net_Sentiment e R.

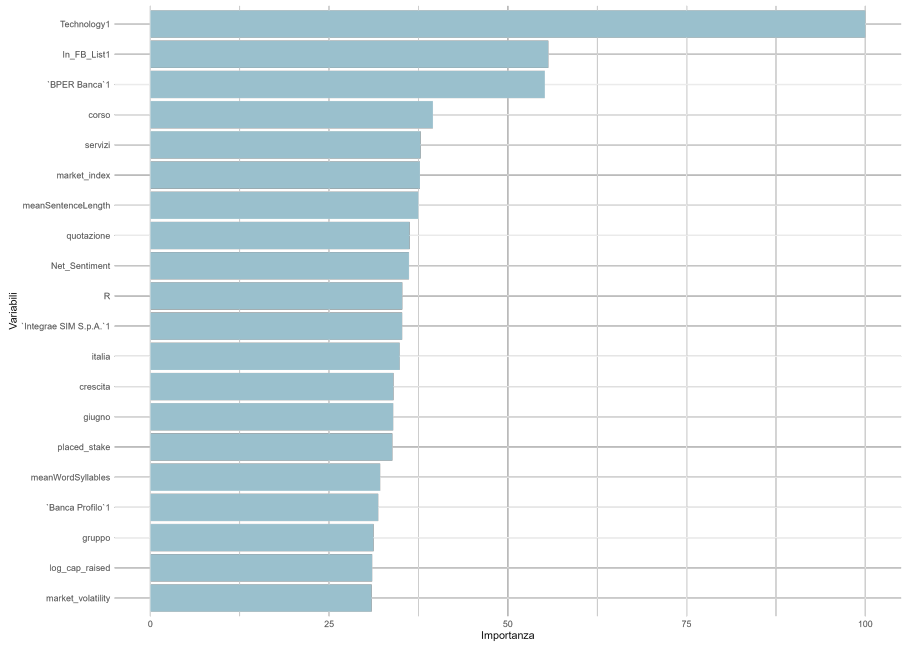

Figura 4.3: Importanza delle 20 variabili più significative nel modello Random Forest.

4.3.3 Support Vector Regression

La griglia dei parametri per il modello SVR con kernel radiale è stata definita con i seguenti parametri di tuning:

- C, rappresenta il parametro di costo e varia da 0.1 a 2 con intervalli di 0.1.

- Sigma, rappresenta il parametro del kernel radiale e varia da 0.005 a 0.1 con intervalli di 0.01.

Il parametro epsilon, che rappresenta il margine di tolleranza, è fissato al valore predefinito di 0.1 dal pacchetto `svmRadial`.
Per l'addestramento del modello, è stato utilizzato il metodo svmRadial, che utilizza il pacchetto `kernelab`. La griglia dei parametri consente di esplorare diverse combinazioni di C e sigma per trovare il modello ottimale.
Il grafico in Figura 4.4 mostra l'andamento dell'indice RMSE rispetto ai parametri C e sigma del modello SVR con kernel radiale e epsilon fissato. Le linee colorate rappresentano alcuni dei diversi valori di sigma, mentre l'asse delle ascisse mostra i valori di C.

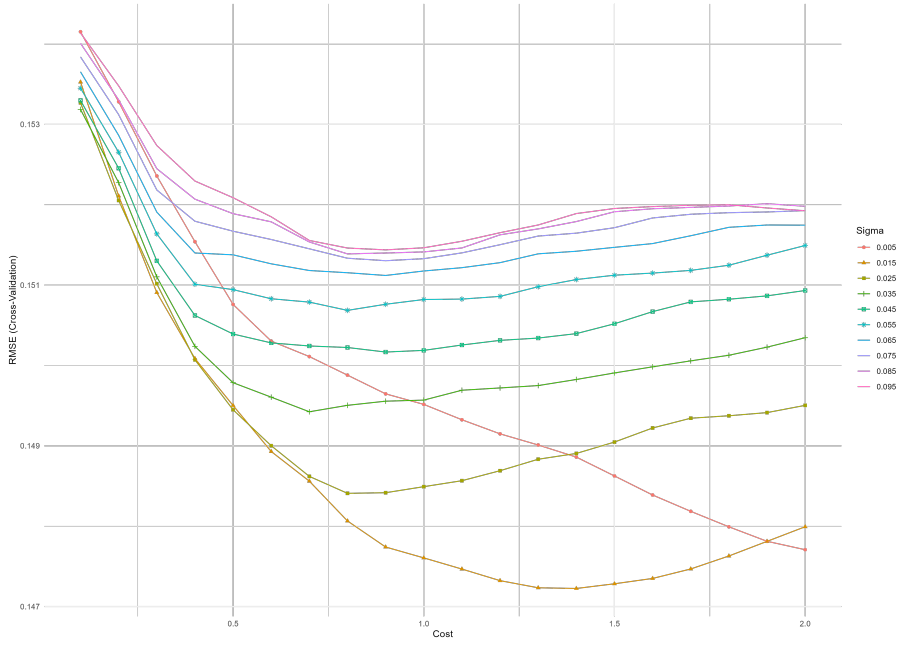

Figura 4.4: Andamento del RMSE in funzione dei parametri C e sigma.

Dal grafico possiamo osservare che al diminuire del parametro sigma e all'aumentare del costo C, l'RMSE in convalida incrociata tende a diminuire. Il valore di

C ottimale è 0.6, dove l'RMSE raggiunge il suo minimo per sigma pari a 0.025.
Il grafico in Figura 4.5 mostra l'importanza delle 20 variabili più significative nel modello SVR con kernel radiale. La variabile più importante risulta nuovamente Technology, confermando ulteriormente i risultati ottenuti fino ad ora. Risultano importanti inoltre le variabili economiche log_cap_raised e market_volatility. Si nota ulteriormente un'importanza associata a variabili legate agli articoli finanziari; risultano infatti importanti sia indicatori come Net_Sentiment e meanSentenceLength sia parole come società, lusso e azioni. In particolare l'importanza associata alla parola lusso risulta coerente con quanto ottenuto nel modello Lasso. All'interno delle 20 variabili più importanti, si notano, come negli altri casi, importanti le variabili associate agli sponsor, quali BPER Banca, Banca Profilo e Banca Popolare di Vicenza e le variabili associate a Facebook, quali In_FB_List e facebook_ipo.

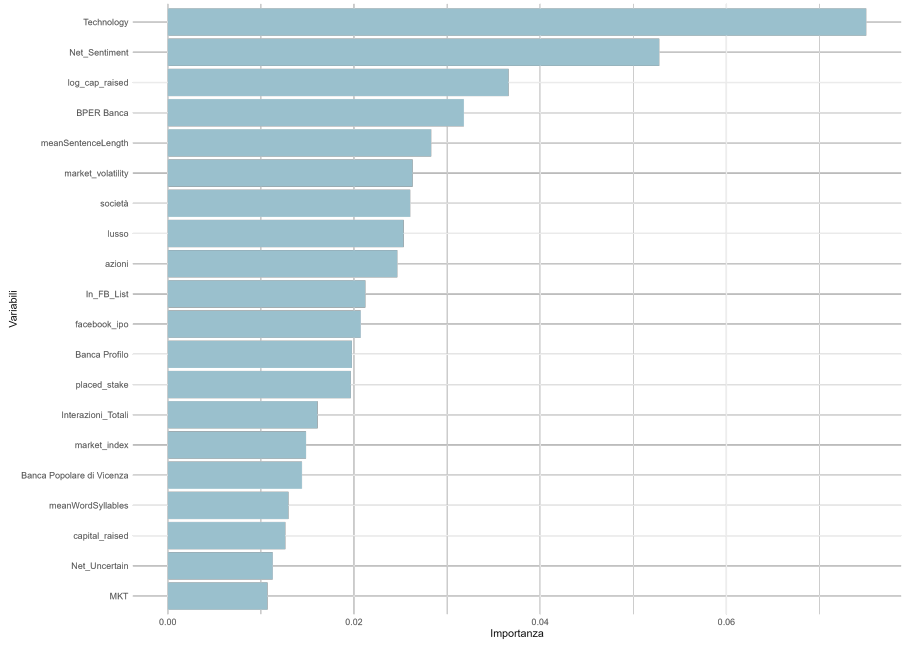

Figura 4.5: Importanza delle 20 variabili più significative nel modello Support Vector Regression.

4.4 Confronto dei modelli

In questa sezione, verranno confrontate le prestazioni dei modelli allenati nella sezione precedente utilizzando il dataset di test. Le metriche di errore utilizzate per valutare le prestazioni dei modelli sono il Root Mean Squared Error (RMSE) e il Mean Absolute Error (MAE). La tabella seguente riassume i valori di queste metriche per ciascun modello:

Modello	**RMSE**	**MAE**
Lasso	0.1834	0.1278
Random Forest	0.1879	0.1309
SVR	0.1929	0.1248

Tabella 4.2: Confronto delle metriche RMSE e MAE per i modelli Lasso, Random Forest e SVR.

Come mostrato nella tabella 4.2, il modello Lasso ha ottenuto il valore di RMSE più basso seguito da Random Forest e infinire SVR. Per quanto riguarda il MAE, l'SVR ha ottenuto il valore più basso, seguito dal Lasso e in conclusione dal Random Forest.

Sebbene il modello Lasso risulti mediamente il miglior modello in termini di RMSE, le differenze tra i tre modelli non sono nette. Questo suggerisce che tutti e tre i modelli possono essere validi strumenti per la previsione dell'underpricing delle IPO, ciascuno con i propri punti di forza. Il Lasso si distingue per la sua capacità di selezione delle variabili, mentre il Random Forest è noto per la sua robustezza e capacità di catturare relazioni non lineari. L'SVR, invece, si distingue per la sua capacità di gestire outlier e dati rumorosi.

In conclusione, la scelta del modello ottimale potrebbe dipendere dalle specifiche esigenze del problema e dalle priorità dell'analisi, quali la necessità di interpretabilità o la capacità di catturare relazioni non lineari.

Capitolo 5

SHAP (SHapley Additive exPlanation) Values

5.1 Revisione della letteratura

L'interpretazione dei modelli complessi di machine learning è fondamentale per garantire fiducia nelle previsioni. Gli SHAP values rappresentano un metodo avanzato per interpretare questi modelli, fornendo spiegazioni coerenti e accurate delle previsioni.

L'ideale per interpretare un modello è utilizzare il modello stesso; tuttavia, per modelli complessi, questa opzione non è praticabile a causa della loro complessità.

Lundberg e Lee (2017) definiscono i metodi di attribuzione delle caratteristiche addittive come funzioni lineari di variabili binarie:

$$g(z^0) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i^0 \tag{5.1}$$

dove $z^0 \in \{0,1\}^M$, M è il numero di caratteristiche di input semplificate e $\phi_i \in \mathbb{R}$. Le caratteristiche di input semplificate mappano ogni caratteristica originale a una

variabile binaria, dove 0 indica l'assenza della caratteristica e 1 indica la sua presenza.
I metodi che rispettano questa definizione attribuiscono un effetto a ciascuna caratteristica, con la somma degli effetti di tutte le attribuzioni che approssima l'output $f(x)$ del modello originale. Tra i metodi che rispettano questa definizione si trovano:

- LIME
- DeepLIFT
- Layer-Wise Relevance Propagation
- Classic Shapley Value Estimation

Una caratteristica fondamentale degli SHAP values è che possiedono tre proprietà desiderabili, che li rendono particolarmente efficaci per spiegare i modelli complessi:

- Accuratezza locale: Quando si approssima il modello originale f per un input specifico x, l'accuratezza locale richiede che il modello di spiegazione corrisponda all'output di f per l'input semplificato x^0, cioè:

$$f(x) = g(x^0) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i^0 \tag{5.2}$$

- Assenza: Se gli input semplificati rappresentano la presenza di una caratteristica, allora l'assenza richiede che le caratteristiche assenti nell'input originale non abbiano alcun impatto alla previsione.
- Consistenza: Se un modello cambia in modo che il contributo di uno specifico input aumenti o rimanga lo stesso indipendentemente dagli altri input, allora l'effetto di quell'input non deve diminuire.

Queste proprietà rendono gli SHAP values un potente strumento per interpretare modelli complessi, poiché forniscono spiegazioni che sono accurate, coerenti e additive.

Tra i metodi che soddisfano queste proprietà, il Kernel SHAP (SHapley Additive exPlanations; Lundberg e Lee, 2017) è ampiamente utilizzato per spiegare le previsioni individuali nei modelli di machine learning. Kernel SHAP sfrutta il concetto di valore di Shapley dalla teoria dei giochi cooperativi per assegnare a ciascuna caratteristica un contributo alla previsione. Esso fa parte della stima classica dei valori di Shapley e calcola i valori di Shapley stimando il valore atteso dell'output del modello considerando tutte le possibili combinazioni di caratteristiche. Lo fa campionando sottoinsiemi di caratteristiche e valutando l'output del modello, combinandoli con medie ponderate.

Quando il modello è non lineare o le feature di input non sono indipendenti, tuttavia, l'ordine in cui le caratteristiche vengono aggiunte al valore atteso conta, e i valori SHAP derivano dalla media dei valori ϕ_i su tutti i possibili ordini.

La Stima Classica dei Valori di Shapley richiede il ricalcolo del modello per tutte le combinazioni possibili di caratteristiche, mentre il Kernel SHAP ottimizza questo processo utilizzando tecniche di campionamento per stimare i contributi delle caratteristiche, mantenendo le stesse proprietà.

5.2 SHAP Values per il modello Random Forest

Il grafico a barre in Figura 5.1 mostra l'impatto medio dei valori SHAP per ciascuna delle 15 variabili con il maggiore impatto nel modello Random Forest.

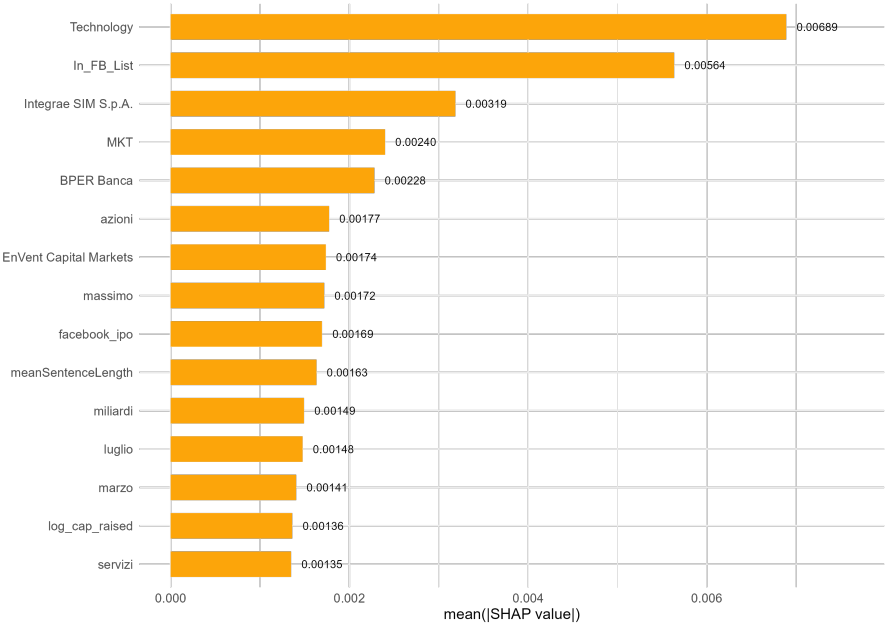

Figura 5.1: SHAP values medi per le variabili più importanti nel modello Random Forest.

Dal grafico possiamo notare che le variabili "Technology", "In_FB_List", e "Integrae SIM S.p.A." hanno i valori SHAP medi più alti, indicando un'influenza significativa sulle previsioni del modello:

- La variabile "Technology" è la più significativa, cambiando la previsione dell'underpricing in media di 0.00689. Questo suggerisce che il settore tecnologico ha un impatto sostanziale sull'underpricing delle IPO

- L'esistenza di una pagina Facebook è la seconda variabile più importante, con una variazione media nella previsione di circa 0.00564.

- La variabile riferita allo sponsor "Integrae SIM S.p.A." mostra un'importanza significativa, cambiando in media la previsione di circa 0.00319.

L'osservazione dei valori SHAP mostra che:

- Gli sponsor risultano importanti, come evidenziato dalla presenza di variabili significative come "BPER Banca", "EnVent Capital Markets", e "Integrae SIM S.p.A.".
- Le variabili legate a Facebook risultano imporanti, come evidenziato da In_FB_List e facebook_ipo.
- Le variabili economiche risultano significative, come dimostrato dall'importanza di "Technology", "MKT" e "log_cap_raised".
- Le variabili testuali sono anch'esse rilevanti, con termini o variabili come "miliardi", "azioni", "massimo", "meanSentenceLength" che influenzano significativamente le previsioni del modello.

Il grafico a sciame d'api in Figura 5.2 mostra la distribuzione dei valori SHAP per ciascuna delle 15 variabili con il maggiore impatto nel modello Random Forest. Il colore nel grafico dei valori SHAP aiuta a comprendere come i valori alti o bassi di ciascuna caratteristica influenzano le previsioni del modello. Il gradiente del colore dal viola al giallo rappresenta valori dal basso all'alto della variabile, la posizione dei punti lungo l'asse x (valore SHAP) indica l'effetto di questi valori sulle previsioni del modello.

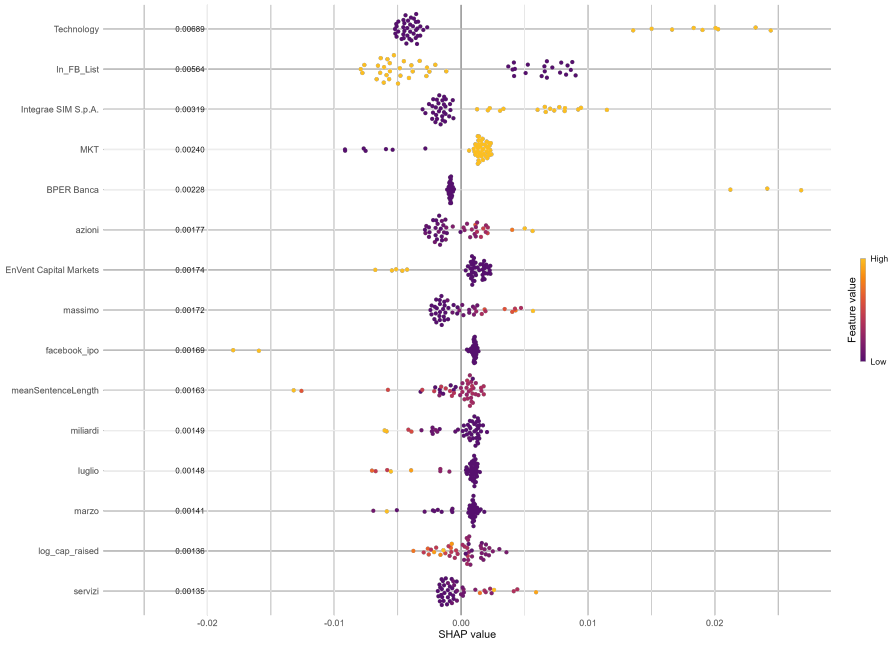

Figura 5.2: Distribuzione dei valori SHAP nel modello Random Forest.

Dal'analisi del grafico 5.2 emergono le seguenti considerazioni:

- Technology: I punti gialli si trovano dopo lo zero, indicando che appartenere al settore tecnologico aumenta l'underpricing delle IPO. I punti viola sono invece prima dello zero, suggerendo che non appartenere al settore tecnologico riduce l'underpricing.

- In_FB_List: Per la variabile "In_FB_List", i punti gialli sono prima dello zero, indicando che disporre di una pagina Facebook diminuisce l'underpricing delle IPO. I punti viola sono dopo lo zero, suggerendo che non disporre di una pagina Facebook aumenta l'underpricing.

- Integrae SIM S.P.A.: La variabile "Integrae SIM S.P.A." mostra punti gialli dopo lo zero, suggerendo che la presenza di questo sponsor aumenta l'underpricing. I punti viola sono prima dello zero, indicando che la mancanza di questo sponsor riduce l'underpricing.

- MKT: La variabile "MKT" mostra punti gialli dopo lo zero, suggerendo che la quotazione nel mercato EGM aumenta l'underpricing. I punti viola sono prima dello zero, indicando che la quotazione nel mercato EXM riduce l'underpricing.

- BPER Banca: Presenta una distribuzione simile a quella di Integrae SIM S.P.A. I punti gialli sono dopo lo zero, indicando che la presenza di questo sponsor aumenta significativamente l'underpricing. I punti viola sono prima dello zero, suggerendo che l'assenza di questo sponsor riduce l'underpricing.

- EnVent Capital Markets: Per la variabile "EnVent Capital Markets", i punti gialli sono prima dello zero, suggerendo che la presenza di questo sponsor riduce l'underpricing delle IPO. I punti viola sono dopo lo zero, indicando che l'assenza di questo sponsor aumenta l'underpricing.

- facebook_ipo: La variabile "facebook_ipo" presenta punti gialli prima dello zero, indicando che citare l'IPO nei post Facebook riduce l'underpricing. I punti viola sono dopo lo zero, suggerendo che non citare l'IPO su Facebook aumenta l'underpricing.

- Per le variabili associate alle parole "azioni", "massimo" e "servizi", possiamo notare che valori bassi delle variabili riducono l'underpricing delle IPO, mentre valori alti aumentano l'underpricing.

- Per le variabili associate alle parole "miliardi", "luglio" e "marzo", possiamo notare che valori bassi della variabile aumentano l'underpricing delle IPO, mentre valori alti diminuiscono l'underpricing.

Per le altre variabili, la distribuzione dei valori SHAP non è chiaramente interpretabile graficamente a causa della sovrapposizione dei punti di colore diverso, che

rende difficile distinguere chiaramente l'impatto delle variabili sulle previsioni del modello.

5.3 SHAP Values per il modello Support Vector Regression

Il grafico a barre in Figura 5.3 mostra l'impatto medio dei valori SHAP per ciascuna delle 15 variabili con il maggiore impatto nel modello SVR.

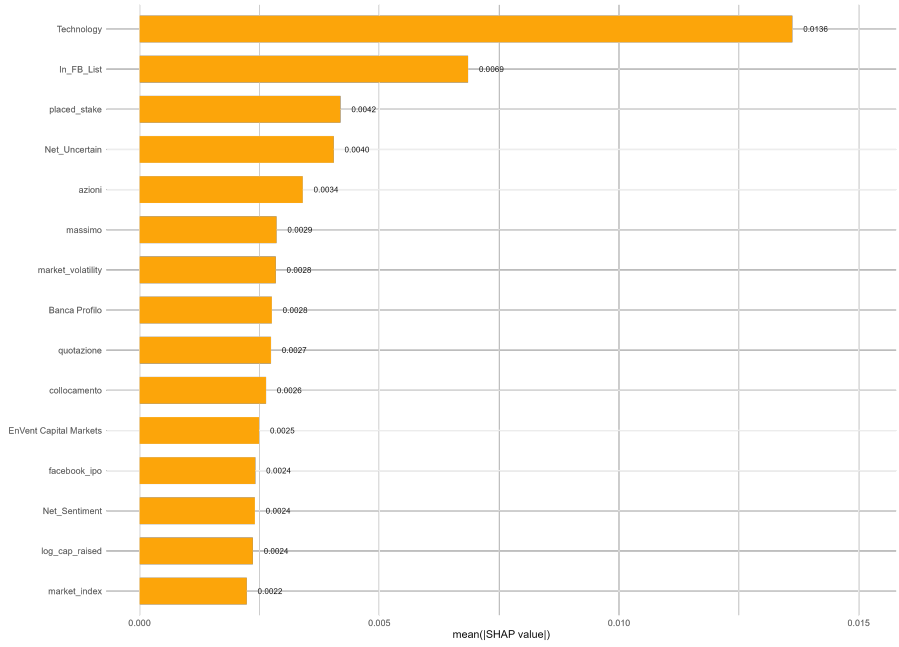

Figura 5.3: SHAP values medi per le variabili con il maggiore impatto nel modello SVR.

Dal grafico possiamo notare che le variabili "Technology" e "In_FB_List" hanno i valori SHAP medi più alti, indicando un'influenza significativa sulle previsioni del modello. In particolare:

- La variabile "Technology" è la più significativa, cambiando la previsione dell'underpricing in media di 0.0136. Questo suggerisce che il settore tecnologico ha un impatto sostanziale sull'underpricing delle IPO.
- L'esistenza di una pagina Facebook è la seconda variabile più importante, con una variazione media nella previsione di circa 0.0069.

L'osservazione dei valori SHAP mostra che:

- Le variabili economiche giocano un ruolo cruciale, come dimostrato dall'importanza associate alle variabili "Technology, "placed_stake", "market_volatility", "log_cap_raised" e "market_index".
- Le variabili legate a Facebook risultano importanti, come evidenziato da "In_FB_List" e "facebook_ipo".
- Le variabili testuali sono anch'esse rilevanti, con termini o variabili come "Net_Uncertain", "Net_Sentiment", "azioni", "massimo", "quotazione" e "collocamento" che influenzano significativamente le previsioni del modello.
- Gli sponsor risultano importanti, come evidenziato dalla presenza di variabili significative come "Banca Profilo" e "EnVent Capital Markets".

Il grafico a sciame d'api in Figura 5.4 mostra la distribuzione dei valori SHAP per ciascuna delle 15 variabili con il maggiore impatto nel modello SVR.

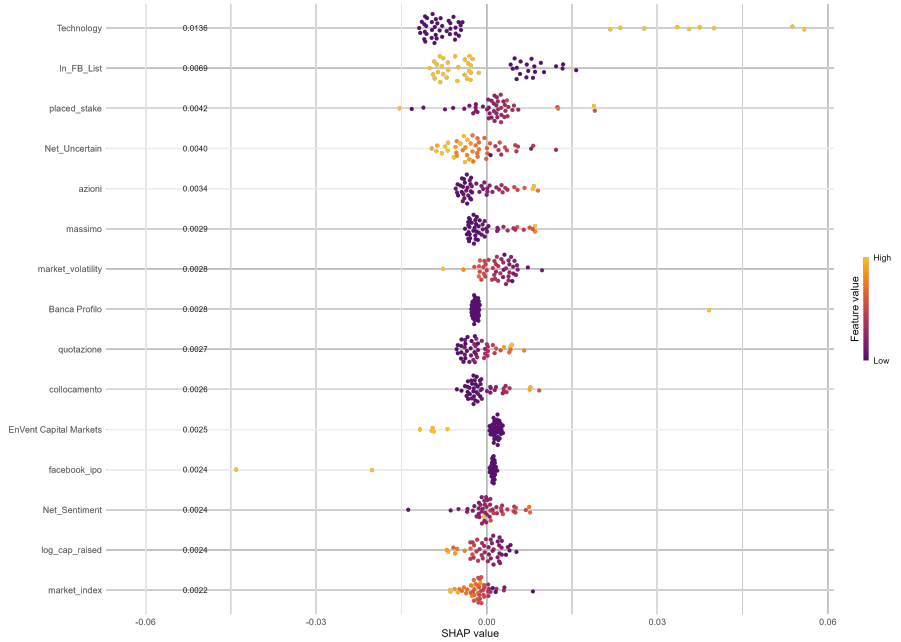

Figura 5.4: Distribuzione dei valori SHAP nel modello SVR.

Dall'analisi del grafico 5.4 emergono le seguenti considerazioni:

- Technology: I punti gialli si trovano dopo lo zero, indicando che appartenere al settore tecnologico aumenta l'underpricing delle IPO. I punti viola sono invece prima dello zero, suggerendo che non appartenere al settore tecnologico riduce l'underpricing.

- In_FB_List: Per la variabile "In_FB_List", i punti gialli sono prima dello zero, indicando che disporre di una pagina Facebook diminuisce l'underpricing delle IPO. I punti viola sono dopo lo zero, suggerendo che non disporre della pagina aumenta l'underpricing.

- Placed_Stake: Sebbene "placed_stake" sia una delle variabili con il maggiore impatto medio SHAP, la distribuzione dei punti gialli sia prima che dopo lo zero rende difficile interpretarne chiaramente l'effetto sull'underpricing.

- Net_Uncertain: Essendo una variabile quantitativa, i punti della variabile "Net_Uncertain" presentano una gradazione di colori meno marcata rispetto alle variabili binarie. I punti più scuri si trovano principalmente dopo lo zero, indicando che un livello basso di incertezza aumenta l'underpricing. I punti gialli sono prima dello zero, suggerendo che un livello elevato di incertezza riduce l'underpricing.

- EnVent Capital Markets: Per la variabile "EnVent Capital Markets", i punti gialli sono prima dello zero, suggerendo che la presenza di questo sponsor riduce l'underpricing delle IPO. I punti viola sono dopo lo zero, indicando che l'assenza di questo sponsor aumenta l'underpricing.

- Per le variabili associate alle parole "azioni", "massimo", "quotazione" e "collocamento" possiamo notare che valori bassi della variabile riducono l'underpricing delle IPO, valori alti invece di questa ne aumentano il valore.

Per le altre variabili, la distribuzione dei valori SHAP non è chiaramente interpretabile graficamente a causa della sovrapposizione dei punti di colore diverso, che rende difficile distinguere chiaramente l'impatto delle variabili sulle previsioni del modello.

Capitolo 6

Risultati e limiti dello studio

6.1 Sintesi dei risultati dei modelli

In questo capitolo verranno riassunti e discussi i risultati ottenuti dai tre modelli utilizzati: Lasso, Random Forest e Support Vector Regression.
L'analisi delle variabili più significative nei diversi modelli hanno evidenziato alcune variabili chiave che hanno un impatto consistente sull'underpricing delle IPO.
La variabile "Technology" emerge come la più rilevante in tutti i modelli analizzati:

- Nel modello Lasso, "Technology" presenta il coefficiente più elevato, indicando un forte impatto positivo sull'underpricing delle IPO.
- Nei modelli Random Forest e SVR, "Technology" presenta i valori SHAP medi più alti rispetto alle altre variabili. L'analisi del grafico a sciame d'api ha inoltre confermato che l'appartenenza al settore tecnologico aumenta i livelli di underpricing.

L'esistenza di una pagina Facebook per un'azienda ("In_FB_List") è un'altra variabile importante:

- Nel modello Lasso, "In_FB_List" ha un coefficiente significativo negativo, indicando che disporre di una pagina Facebook tende a ridurre l'underpricing.

- Nei modelli Random Forest e SVR, "In_FB_List" ha valori SHAP medi elevati, confermando la sua importanza nella riduzione dell'underpricing, come confermato anche dai due grafici a sciame d'api.

Il risultato ottenuto dall'analisi di questa variabile risulta controintuitivo. Come riportato nella prima parte di letteratura, nello specifico nell'articolo di Lundamark et al. (2017), l'esistenza di una pagina social dell'azienda dovrebbe condurre ad avere livelli di underpricing superiori. Tuttavia, i risultati ottenuti possono essere limitati dal fatto che molte pagine non sono molto utilizzate. Per molte aziende infatti, non è stato possibile scaricare i post a causa della mancanza di seguito e del tracciamento da parte di CrowdTangle. I limiti di questa variabile sono inoltre evidenziati anche dal fatto che, anche per quelle aziende per cui è stato possibile scaricare i post, poche parlano del proprio processo di quotazione in borsa sui social, come indicato dalla frequenza della variabile "facebook_ipo".
Gli sponsor risultano essere variabili cruciali per l'underpricing delle IPO:

- Nel modello Lasso, sponsor come "Alantra Capital Markets", "Banca Profilo" e "BPER Banca" mostrano coefficienti positivi significativi, "Banca Popolare di Vicenza" coefficiente negativo significativo.

- Nei modelli Random Forest e SVR, sponsor quali "BPER Banca", "Integrae SIM S.p.A.", "Envent Capital Markets" e "Banca Profilo" presentano valori SHAP medi elevati, suggerendo un impatto rilevante sull'underpricing.

Visto l'impatto significativo di alcuni sponsor sulle previsioni dell'underpricing, risulta ragionevole ipotizzare che il ruolo di garante del profilo qualitativo dell'emittente svolto dagli sponsor può avere un effetto diretto sulla fiducia degli investitori e, di conseguenza, sull'underpricing delle IPO.
Le variabili testuali risultano anch'esse rilevanti:

- Nel modello Lasso, il termine "lusso" presenta un coefficiente significativo positivo e il termine "miliardi" coefficiente negativo significativo.
- Nei modelli Random Forest e SVR, variabili come "Net_Uncertain", "Net_Sentiment", "meanSentenceLength" e le variabili associate ai vari termini presentano valori SHAP medi elevati.

Questi risultati evidenziano l'importanza di considerare anche le informazioni derivate dai testi, per ottenere una comprensione più completa dei fattori che influenzano l'underpricing delle IPO.
Oltre alla variabile "Technology" già citata, le variabili economiche mostrano una rilevanza limitata:

- Nel modello Lasso, nessuna variabile economica viene selezionata.
- Nei modelli Random Forest e SVR, le variabili "MKT", "log_cap_raised", "placed_stake", "market_volatility" e "market_index" mostrano valori SHAP medi elevati, suggerendo una certa rilevanza.

Tuttavia, è importante notare che, sebbene queste variabili economiche presentino valori SHAP medi elevati nei modelli Random Forest e SVR, non sono mai tra le variabili con l'apporto maggiore. Questo suggerisce che, al di là della variabile "Technology", le altre variabili economiche non giocano un ruolo importante nell'underpricing delle IPO. L'assenza di selezione di variabili economiche nel modello Lasso rafforza ulteriormente questa osservazione. Già dall'analisi esplorativa si era

evidenziato che queste variabili non avevano un impatto significativo sulla risposta. In conclusione, è possibile affermare che:

- L'analisi effettuata mostra che l'esistenza di una pagina Facebook ha un impatto significativo sull'underpricing, sebbene in direzione negativa. Non sono risultate significative le altre variabili legate ai social media.
- Le variabili testuali legate al sentiment e all'incertezza hanno mostrato una certa rilevanza nei modelli Random Forest e SVR. Questo conferma che un sentiment positivo nei media può contribuire ad aumentare l'underpricing delle IPO.
- La variabile legata al numero di articoli non è risultata significativa in nessuno dei modelli analizzati.

Questi risultati confermano parzialmente le ipotesi iniziali presentate nella sezione 1.3.

6.2 Limiti dello studio

Nonostante alcuni risultati significativi, lo studio presenta limiti che devono essere considerati:

- Ampiezza del campione: I dati utilizzati per questo studio arrivano fino all'inizio del 2022. Tuttavia, il periodo tra il 2019 e il 2021 è stato caratterizzato da un "hot market" con un numero significativo di aziende che si sono quotate in borsa. Questo trend è continuato anche negli anni successivi. Integrare i dati delle IPO avvenute in questo periodo recente potrebbe aumentare la numerosità del campione e migliorare la qualità dei dati social. Le aziende che si quotano in borsa oggi utilizzano i social media in modo molto diverso rispetto

a quindici anni fa, con strategie di comunicazione e marketing più sofisticate e mirate, che riflettono meglio le aspettative moderne degli investitori.

- Dati relativi a Facebook limitati: Sebbene Facebook sia stato utilizzato come fonte di dati social, potrebbe non essere il social network più rilevante in ambito finanziario. Numerosi studi suggeriscono che Twitter (ora X) ha un impatto maggiore nel contesto finanziario. Tuttavia, le recenti modifiche alla piattaforma, inclusa la restrizione al download dei post, non hanno reso possibile utilizzare questa fonte di dati. Questo rappresenta un limite significativo, poiché i dati di X avrebbero potuto fornire un quadro più accurato dell'interesse e del sentiment del mercato.

- Utilizzo limitato di media: Lo studio ha effettuato il web scraping solo dai siti di MilanoFinanza e SoldiOnline. Sebbene queste siano fonti importanti, escludere altre testate potrebbe aver limitato la diversità delle informazioni raccolte. L'inclusione di dati provenienti da una varietà più ampia di fonti di notizie potrebbe fornire un quadro più completo dell'interesse e del sentiment del mercato.

- Sebbene le variabili economiche come i prezzi minimo e massimo dell'IPO, che rappresentano l'intervallo di prezzo all'interno del quale viene individuato il prezzo finale di sottoscrizione dei titoli, siano disponibili nel dataset, la presenza di molti valori mancanti ha limitato il loro utilizzo nell'analisi. La raccolta completa di questi valori è complicata e richiede tempo, ma potrebbe offrire una comprensione più dettagliata delle dinamiche di prezzo nelle IPO.

- La variabile "picco", derivata da Google Trends, indica la presenza di un aumento delle ricerche su Google per ciascuna azienda nei quindici giorni precedenti la data di quotazione in borsa. Tuttavia, la variabile presenta delle

limitazioni importanti. In molti casi, la funzione gtrends non è stata in grado di scaricare i dati a causa di un numero insufficiente di query di ricerca per il termine specifico. Un possibile miglioramento potrebbe essere considerare un periodo di ricerca più ampio per aumentare la disponibilità dei dati, anche se ciò potrebbe ridurre la capacità di rilevare un interesse specifico pre-quotazione. L'obiettivo principale resta quello di determinare l'attenzione pre-quotazione in borsa, quindi l'approccio andrebbe calibrato per bilanciare disponibilità dei dati e precisione dell'analisi.

Questi limiti devono essere tenuti in considerazione quando si interpretano i risultati di questo studio e si considerano possibili applicazioni future.

Capitolo 7

Conclusioni

La tesi ha esaminato l'influenza che fattori da un lato economici e dall'altro legati a media, social media e web hanno sulle performance delle aziende. Si è proposto, quindi, un modello di analisi che combina dati economici tradizionali con nuove fonti di dati digitali, e lo si è applicato all'orizzonte temporale 2010-2022 e al contesto italiano.

I risultati hanno mostrato che l'impatto più significativo per l'underpricing delle IPO è rappresentato dall'appartenenza al settore tecnologico. Inoltre, si è riscontrato che anche lo sponsor ha un ruolo rilevante, e che quindi il profilo qualitativo di chi sostiene l'azienda ha un'influenza sul mercato azionario. Si è constatato, invece, che le variabili legate sia ai media tradizionali che ai social media hanno un impatto più marginale. Come ribadito sopra, i risultati ottenuti si limitano a un arco temporale specifico. Si può presumere che, in archi temporali futuri, alcuni dei fattori analizzati, come ad esempio i social media, possano avere impatti diversi sulle aziende, e che quindi l'influenza che hanno sui mercati finanziari possa variare. Le future ricerche dovrebbero quindi includere queste variabili, in modo da approfondire le dinamiche che caratterizzano il processo di quotazione in borsa e poter migliorare la completezza dell'analisi.

Bibliografia

[1] Debasish Basak, Srimanta Pal, Dipak Chandra Patranabis et al. "Support vector regression". In: *Neural Information Processing-Letters and Reviews* 11.10 (2007), pp. 203–224.

[2] Alan L Beller, Tsunemasa Terai e Richard M Levine. "Looks can be deceiving: A comparison of initial public offering procedures under Japanese and US securities laws". In: *Law and Contemporary Problems* 55.4 (1992), pp. 77–118.

[3] Gérard Biau e Erwan Scornet. "A random forest guided tour". In: *Test* 25 (2016), pp. 197–227.

[4] Andrea Ceron, Luigi Curini e Stefano Maria Iacus. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. Vol. 9. Springer Science & Business Media, 2014.

[5] Enrico Maria Cervellati, Antonio Carlo Francesco Della Bina, Pierpaolo Pattitoni et al. "The efficiency of the Italian stock exchange: market reaction following changes in recommendations". In: *Corporate, Ownership & Control Journal* 5.2 (2008), pp. 432–48.

[6] Ansley Chua. "Market conditions, underwriter reputation and first day return of IPOs". In: *Journal of Financial Markets* 19 (2014), pp. 131–153.

[7] Jonathan Clarke, Tomas Jandik e Gershon Mandelker. “The efficient markets hypothesis”. In: *Expert financial planning: Advice from industry leaders* 7.3/4 (2001), pp. 126–141.

[10] Vikas Gupta, Shveta Singh e Surendra S Yadav. “The impact of media sentiments on IPO underpricing”. In: *Journal of Asia business studies* 16.5 (2022), pp. 786–801.

[11] Patricia J Hughes e Anjan V Thakor. “Litigation risk, intermediation, and the underpricing of initial public offerings”. In: *The Review of Financial Studies* 5.4 (1992), pp. 709–742.

[13] Kentaro Ito e Ryohei Nakano. “Optimizing support vector regression hyperparameters based on cross-validation”. In: *Proceedings of the International Joint Conference on Neural Networks, 2003.* Vol. 3. IEEE. 2003, pp. 2077–2082.

[14] Narasimhan Jegadeesh, Mark Weinstein e Ivo Welch. “An empirical investigation of IPO returns and subsequent equity offerings”. In: *Journal of Financial Economics* 34.2 (1993), pp. 153–175.

[15] Daniel Kahneman e Amos Tversky. “Prospect Theory: An Analysis of Decision under Risk”. In: vol. 47. 2. [Wiley, Econometric Society], 1979, pp. 263–291.

[16] Matti Keloharju. “The winner’s curse, legal liability, and the long-run price performance of initial public offerings in Finland”. In: *Journal of Financial Economics* 34.2 (1993), pp. 251–277.

[17] Laura Xiaolei Liu, Ann E Sherman e Yong Zhang. “An attention model of IPO underpricing, with evidence on media coverage”. In: *Unpublished Working Paper, De Paul University* (2014).

[18] Laura Xiaolei Liu, Ann E Sherman e Yong Zhang. “The long-run role of the media: Evidence from initial public offerings”. In: *Management Science* 60.8 (2014), pp. 1945–1964.

[19] Tim Loughran e Bill McDonald. “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks”. In: *The Journal of finance* 66.1 (2011), pp. 35–65.

[20] Scott M Lundberg e Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).

[21] Leif W Lundmark, Chong Oh e J Cameron Verhaal. “A little Birdie told me: Social media, organizational legitimacy, and underpricing in initial public offerings”. In: *Information Systems Frontiers* 19 (2017), pp. 1407–1422.

[22] Philipp Probst, Marvin N Wright e Anne-Laure Boulesteix. “Hyperparameters and tuning strategies for random forest”. In: *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9.3 (2019), e1301.

[23] Jay R Ritter. “Behavioral finance”. In: *Pacific-Basin finance journal* 11.4 (2003), pp. 429–437.

[24] Jay R Ritter e Ivo Welch. “A review of IPO activity, pricing, and allocations”. In: *The journal of Finance* 57.4 (2002), pp. 1795–1828.

[25] Stuart Rose et al. “Automatic keyword extraction from individual documents”. In: *Text mining: applications and theory* (2010), pp. 1–20.

[26] Andrea Sciandra e Riccardo Ferretti. “Predictive performance comparisons of different feature extraction methods in a financial column corpus”. In: *SIS 2022-Book of Short Papers* (2022), pp. 421–427.

[28] Timm O Sprenger et al. “Tweets and trades: The information content of stock microblogs”. In: *European Financial Management* 20.5 (2014), pp. 926–957.

[29] CrowdTangle Team. *CrowdTangle*. [1826461]. Menlo Park, California, United States: Facebook, 2024.

[30] Emanuele Teti e Ilaria Montefusco. "Corporate governance and IPO underpricing: evidence from the italian market". In: *Journal of Management and Governance* 26.3 (2022), pp. 851–889.

[31] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

Sitografia

[8] Milano Finanza. *Milano Finanza*. 2024. URL: https://www.milanofinanza.it/.

[9] Google Trends. *Google Trends*. 2024. URL: https://trends.google.it/trends/.

[12] Borsa Italiana. *Borsa Italiana*. 2024. URL: https://www.borsaitaliana.it/.

[27] SoldiOnline. *SoldiOnline*. 2024. URL: https://www.soldionline.it/.