Progetto Data Mining

Marco Gelsomini

2022-08-25

Abstract

In questo progetto si vuole costruire un modello predittivo per stabilire se le performance di un titolo azionario saranno positive o negative a alla fine dell' anno.
Tutte le società quotate al NYSE sono tenute a compilare e rendere pubblico un form (*form 9k*) richiesto dalla SEC (*Security Exchange Commission*) nel quale vengono indicati svariati indicatori di bilancio. Ciò permette agli investitori di conoscere lo stato di salute delle imprese.
Si utilizzeranno tali indici come esplicative per le performance dei prezzi dei titolo nell' anno successivo alla pubblicazione di questi indicatori. In questa analisi dopo aver manipolato il dataset sulla base di una piccola conoscenza di dominio si è proceduto a valutare la capacità di classificazione del modello di regressione logistica (binary) e del metodo dei nearest neighbor.
Dopo aver affiancato ai due modelli alcuni metodi di pre processing e feature engineering si è scelto come miglior modello la regressione logistica supportata da ribilanciamento della classe risposta dei dati.

1 Introduzione

Nel mondo degli investimenti spesso si cerca una strategia che permetta di "battere il mercato". Ciò significa avere un ritorno sull' investimento maggiore rispetto a quello offerto dai principali indici globali (*investimento attivo*).
Tra i tanti metodi di investimento si ritrova anche il cosidetto *value investing*. Cioè si analizza il quadro economico e finanziario delle società al fine di scovare quelle che, secondo l' analisi, sono sottovalutate in termini di prezzo assegnato loro dai mercati finanziari in un dato momento.
In questo progetto si cerca di applicare un' analisi statistica su dati del passato che legano indici e metriche delle aziende alla loro performance in borsa nell' anno successivo alla pubblicazione di tali indici.
In altri termini si cerca di utilizzare metodi statistici come classificazioni e/o regressioni per scoprire a posteriori i legami tra stato di salute dichiarato e movimento futuro delle quotazioni.

2 Materiale e metodi

2.1 Materiali

Il dataset è messo a disposizione sulla piattaforma Kaggle. Essendo una risorsa esterna già predisposta non sono noti completamenti tutti i passaggi di codifica, costruzione e integrazione dei dati.
Il dataframe fa riferimento ai dati pubblicati nei report a fine 2018 e, nella sua forma originale, ha 224 variabili per 4392 aziende quotate al NYSE.
221 sono variabili numeriche, 1 variabile di categoria riguarda il settore di attività e infine si ha a disposizione la variabile risposta con la performance del titolo sul mercato da inizio 2019 a fine 2019. E' già presente anche la versione dicotomizzata(performance positiva=1, negativa=0).

2.1.1 Tipologia di variabili

Un primo aspetto che si nota è la presenza di due tipi di variabili:

1. **Variabili assolute**: registrano metriche espresse tipicamente in valore monetario e assumono valori sulla scale delle migliaia come ad esempio fatturato, profitto lordo, spese operative ecc.

2. **Variabili relative**: sono tipicamente costruzioni di indici partendo dalle variabili di cui sopra (sono chiamate anche anche *multipli* o *ratio*). Come ad esempio il margine di profitto, il price earning ratio, il return on equity ecc.
   Quest' ultime permettono un confronto più omogeneo tra aziende con capitalizzazione e/o dimensioni differenti. E' plausibile pensare che il confronto sia ancora più sensato quando questi ratio vengono analizzati per aziende che operano nello stesso settore.

2.1.2 Presenza di zeri

Un' altro aspetto che emerge esplorando il dataset è la presenza di zeri. Per semplicità e non avendo la conoscenza di dominio necessaria si assume che siano effettivamenti valori numerici pari a zero e non siano per esempio stati compilati e pensati come pseudo missing values.
Nel dataset di partenza ci sono il 13.58% di zeri sull' intero dataframe.
Nella figura 1 si mostrano la percentuali di zeri per riga e per variabile.

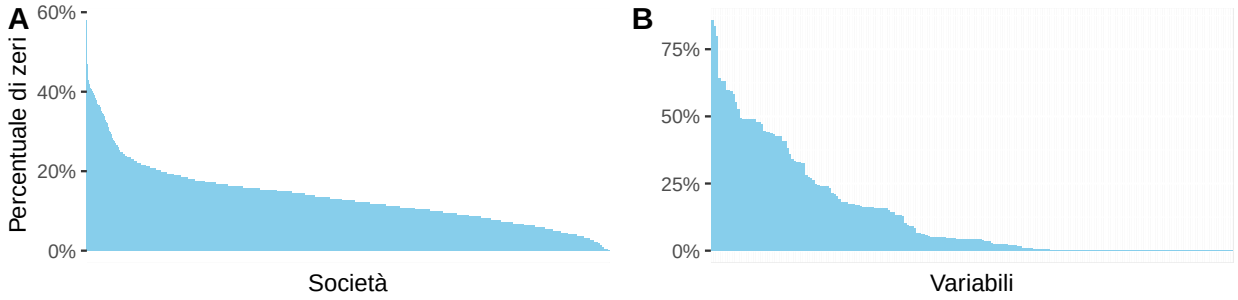

Figure 1: Percentuale di zero per riga e per colonna

Emerge un ristretto gruppo di società che ha compilato con molti zeri (sopra al 20%).
Si nota infine tre variabili con una percentuale di zeri sopra il 75%.

2.1.3 Presenza di missing values

Per quanto riguarda i campi che contengono dei missing values si osserva che nel dataframe di partenza la loro percentuale sul totale è del 10.02%.
Nella figura 2 si mostrano la percentuali di NA per riga e per variabile.

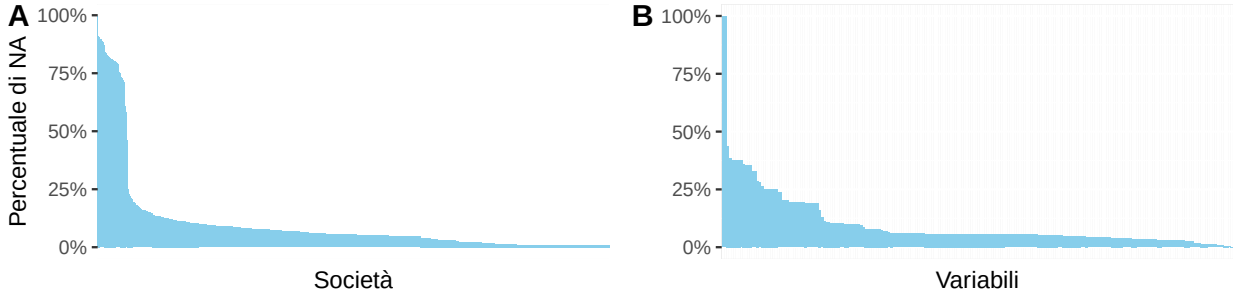

Figure 2: Percentuale di NA per riga e per colonna

Si nota un gruppo di aziende che registra un percentuale visibilmente maggiore rispetto al resto. Mentre le variabili estreme, in questo senso, sono solamente due.

Sembra lecito chiedersi se la presenza di omissioni nella compilazione da parte delle azienda abbia un' influenza sulla loro performance in borsa.
Il grafico seguente indaga, almeno visivamente, questa relazione.

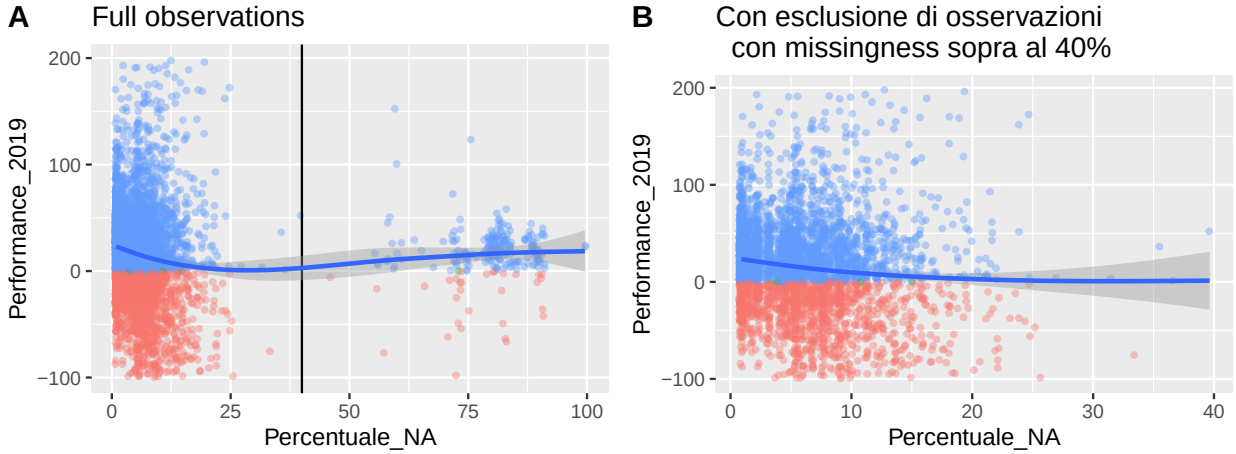

Figure 3: Relazione Missingness e variabile risposta

La nuvola di punti non sembra seguire un pattern particolare dettato dalla percentuale di NA.

2.1.4 Class imbalance

Un altro aspetto importante da indagare è la proporzione dei due valori della variabile risposta. Notariamente nell' anno 2019 si è verificato un rialzo diffuso su quasi tutto il mercato azionario.
Ciò emerge anche in questo dataframe: - Valore 1, performance positiva: 30.6%.
- Valore 0, performance positiva: 69.4%.
Si registra quindi un prevalanza doppia delle aziende con classe 1.

2.1.5 Distribuzioni

E' utile avere un'idea delle distribuzioni delle variabili.
Nella figura 4 si mostra l' indice di kurtosi per ogni variabile con distinzione sulle due classi.

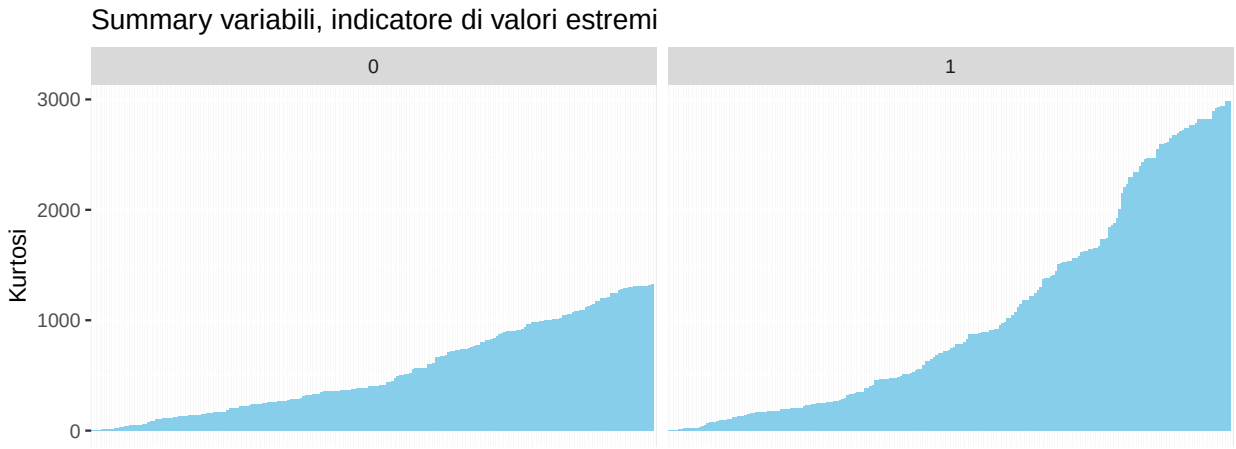

Figure 4: Kurtosi per ogni variabile con distinzione tra classi

Ricordiamo la definizione di kurtosi: $E\left[(\frac{X-\mu}{\sigma})^4\right]$.
In generale si evidenzia una lontananza di tutte le variabili dalla distribuzione normale. Le variabili infatti presentano code più pesanti e quindi maggior probabilià di valori estremi.
Si nota inoltre che per le performance positive (grafico a destra) il livello generale di valori estremi è maggiore. Ciò, in realtà, potrebbe derivare semplicemente dal fatto che ci sono più osservazione per performance dei titoli positive e quindi più facile registrare valori estremi.
In ogni caso va tenuto presente in caso di utilizzo di modelli e metodi che si basano su assunzione di normalità delle features.

2.1.6 Correlazioni e variabili ridondanti

Infine esplorando le variabili a disposizione si nota come in realtà ci siano diverse variabili doppiate.
Potrebbe essere stato un errore in fase di codifica. Per esempio si trova *PriceSalesRatio* e *PriceToSalesRatio*.
Comunque sarà necessario rimuoverle.

2.2 Metodi

2.2.1 Pre processing

Avendo osservando alcune criticità dei dati a disposizione occorre adottare alcune azioni di pre processing:

2.2.1.1 Solo variabili relative Per prima cosa, dato l' elevato numero di variabili, si decide di utilizzare solo quelle appartenti alla categoria dei ratio/moltiplicatori. La ragione è che, come accennato in precedenza, gli investitori, tipicamente, non fanno un confronto tra aziende usando variabili prettamente assolute e monetarie. Infatti esse sono influenzate da "effetti di scala" che dipendono ad esempio dalla grandezza dell' azienda ma non catturano qualità o capacità di generare profitti della società. Dopo questo passaggio il numero il numero di features scende a 108.

2.2.1.2 Rimozione variabili ridondanti In secondo luogo, tra le features rimanenti, vanno rimosse quelle che descrivono nei fatti la stessa metrica.
Dopo questo passaggio il numero di colonne si riduce a 83.

2.2.1.3 Rimozione missingness estrema Per quanto riguarda i valori mancanti si è deciso di rimuovere le osservazione con una proporzione di missing visivamente maggiore rispetto al resto. Lo stesso approccio lo si utilizza per la variabili.
In figura 5 si mostrano righe e colonne con troppi missing values dopo avere ridotto in parte il numero di variabili con i passaggi di cui sopra.
Si decide quindi, in modo arbitario ma visualmente logico, di inserire un treshold per la righe del 25% e per le colonne del 50%.

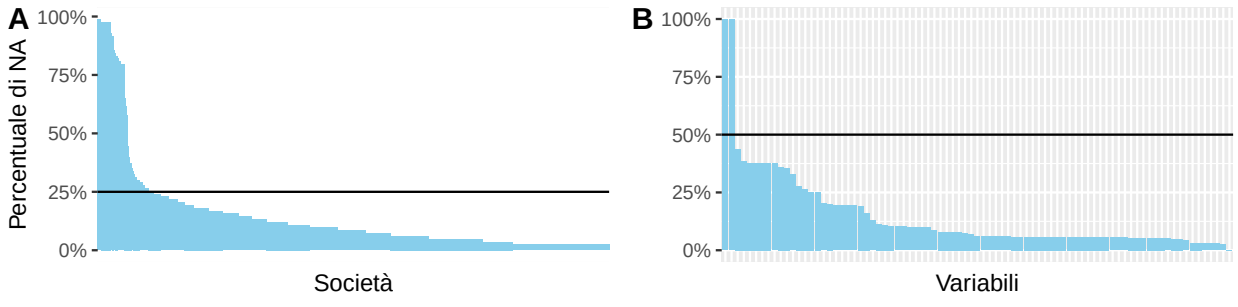

Figure 5: Percentuale di NA per riga e per colonna

Dopo queste operazioni il dataframe ripulito è costituito da 3915 righe e 81 colonne ha una percentuale di missing values del 6.58% e una percentuale di zeri del 13.83%.

La proporzione di classe “performance negativa” è del 30.8% mentre quella positiva è pari al 69.2%. La restante quota di missing values verrà poi imputata in fase di stima.

2.2.2 Breve analisi esplorativa

Prima di provare ad applicare un modello si cercano a livello grafico realazioni tra le features e la variabile risposta. Nel grafico seguente si mostrano le performance dei titoli in base al settore di appartenenza.

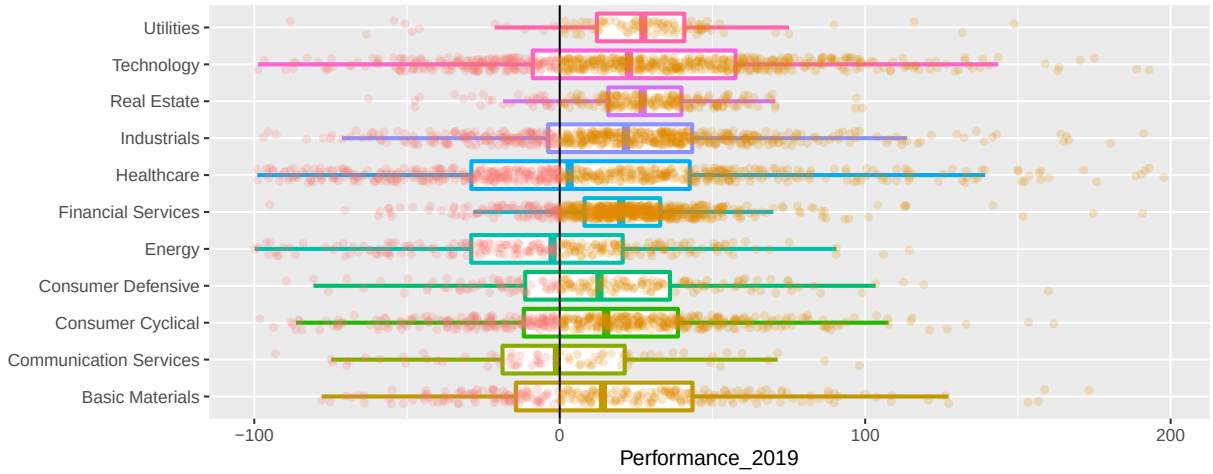

Figure 6: Distribuzione variabile risposta per settore

Il settore sembra essere un buon predittore per le performance.
Essendoci ancora molte variabili, nel grafico in figura 7 si sintentizzano tutte in base alle mediane condizionate sia alla classe di performance sia al settore di attività. Vengono riportate solo le prime 5 features, per ogni settore, che registrano la più grande differenza tra performance positive e negative.
Si ricorda che il settore di attività è un fattore importante quando si confrontano aziende nel value investing. Si utilizzano le mediane per evitare influenze da parte di valori anomali.

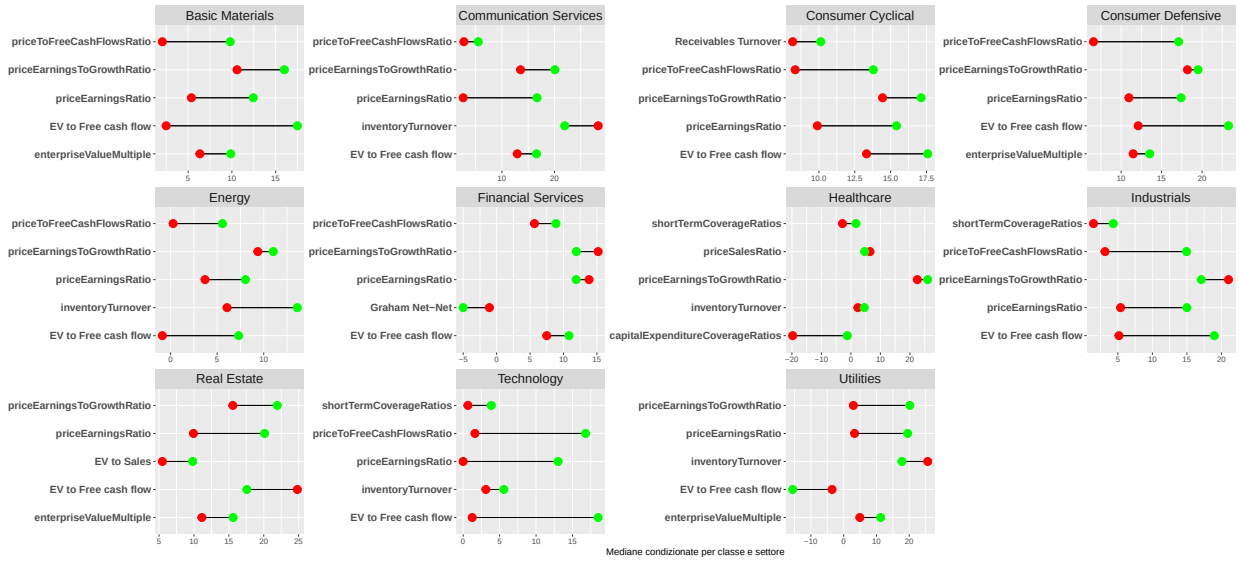

Figure 7: Mediane condizionate per classe e per settore,pallini rossi performance negative, verdi positive

Effettivamente per settori diversi ricadono in questa sorta di classifica metriche diverse. Si possono però

individuare alcune variabili ricorrenti. Per esempio *EV to Free cash flow*, *priceEarningsRatio* oppure *priceEarningsToGrowthRatio*.

2.2.3 Modelling e feature enginering

Dopo questa prima fase di pulizia ed esplorazione dei dati si procede ad applicare alcuni modelli in combinazione con alcuni metodi e operazioni di feature enginering. Si cerca quindi di trovare la combinazione che ottiene la miglior performance.
Si sceglie di vagliare la regressione logistica e il metodo dei nearest neighbor.
In primo luogo si stabilisce la migliore combinazione tra modello e sets di manipolazioni delle features.
Il modello migliore, in combinazione con il suo pro preprocessing, verrà poi testato su dati del test, ovvero su dati mai visti in fase di training per valutare se c'è stato overfitting in fase di stima e tuning dei parametri.

2.2.3.1 Modelli usati

- La **regressione logistica** è un modello di regressione che rientra nella famiglia dei GLM(Generalized Linear Model). In particolare in caso di variabile risposta binaria a cui si associano caratteristiche (features) non ripetute lungo le osservazioni si parla di *Binary Regression*.
  La specificazione del modello prevede quindi la seguente equazione: $logit(\theta) = \eta$ dove θ indica la probabilità di performance positiva del titolo mentre η indica il predittore lineare ovvero la combinazione lineare di parametri (che verrano stimati) e di valori rispettivi delle covariate. La funzione $logit(\theta)$ è definita come: $log(\frac{\theta}{1-\theta})$. Ciò permette di riportare le probabilità di un evento sulla scala dei reali. Oltre ad altri vantaggi computazionali in fase di stima.

- Il metodo dei **nearest neighbor** invece è un modello non parametrico in quanto non avviene una vera e propria stima di parametri in fase di training. Non richiede particolari assunzioni sulla distribuzione delle variabili.
  Questo modello dato una osservazione in analisi x_0 e un numero di vicini K identifica un gruppo Ω_0 di K osservazioni più vicine a x_0 sulla base di una metrica di distanza. Dopodichè calcola le probabilità condizionate di appartenere ad ogni classe della risposta per l' osservazione x_o come segue: $Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i\in\Omega_0} I(y_i = j)$. Quindi assegnerà per osservazione x_o la classe con la probabilità maggiore.
  La fase di training perciò servirà solamente per trovare il numero K ottimale.

2.2.3.2 Pre processing- feature engineering Per quanto riguarda il preprocessing si è scelto di confrontare 2 insieme di procedure o "ricette":

1. Variabile risposta spiegata da tutte le metriche presenti alla fine del processo di pulizia.
   Imputazione dei missing vlaues tramite mediane per evitare influenze di valori anomali.
   Rimozioni di variabili con varianza nulla. E standardizzazione di tutte le variabli numeriche.
2. Come ricetta numero 1 con in aggiunta:

- Filtro per rimuovere variabili corralate. In particolare si fissa una soglia tra 0 e 1 e un algoritmo cercherà di rimuovere il numero minimo di variabili in modo tale da avere, in valore assoluto, tutte le correlazioni tra variabili sotto questa soglia. Il livello ottimo verrà trovato tramite cross validation.
- Campionamento ripetuto aggiuntivo per osservazioni con la classe minoritaria. Anche qui il parametro che stabilisce quante osservazioni reimmettere nel training set sarà trovato tramite cross validation. Se tale parametro ad esempio è 1 significa estrarre casualmente osservazioni con classe minoritaria in modo da avere la stessa proporzione tra le due classi nei dati.

2.3 Risultati

Di seguito in figura 8 vengono mostrati i risultati delle combinazioni tra modelli e ricette in base alla metrica dell' accuracy. I modelli rappresetati sono quelli con i parametri ottimizzati. Per esempio il modello KNN con la ricetta numero 1 ha come numero di vicini $k = 16$. Mentre per la ricetta 2 ha ottimizzato un valore di $k = 20$.

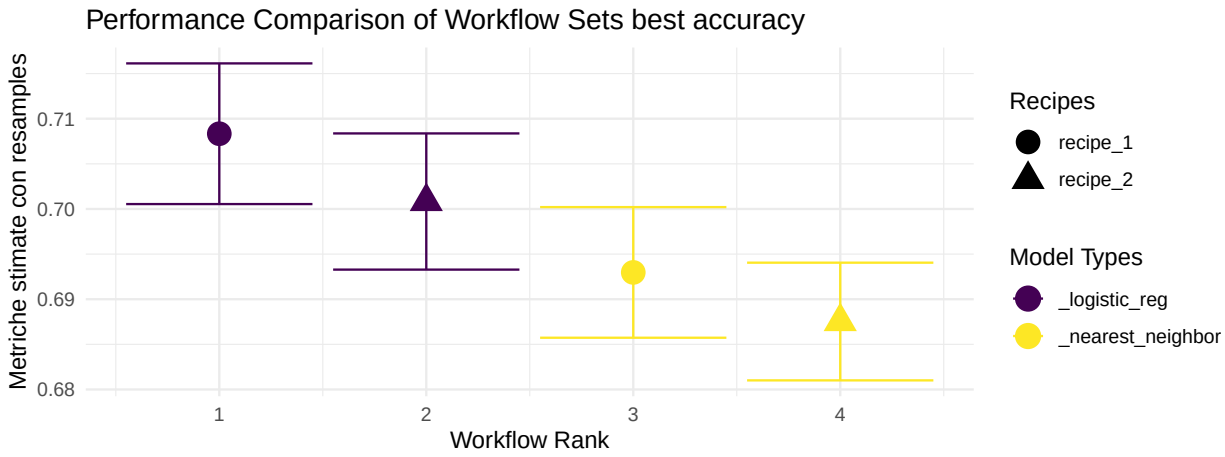

Figure 8: Migliore accuracy tra modelli e preprocessing- Training set

La migliore accuracy nei dati di test la ottiene la regressione logistica con la ricetta numero 1. In questa combinazione non è stato fatta nessuna ottimizzazione e non è stata risolta la questione della class imbalance.

Se però osserviamo (fiugura 9) i punteggi relativi alla sensitività (in fase di stima il successo è stato codificato come classe 0 cioè performance negativa, sarebbe più sensato avere specificità che cattura classe minoritaria) ovvero alla capacità di predirre aziende che hanno avuto performance negative il modello base risulta essere il peggiore con un valore di 0.28.

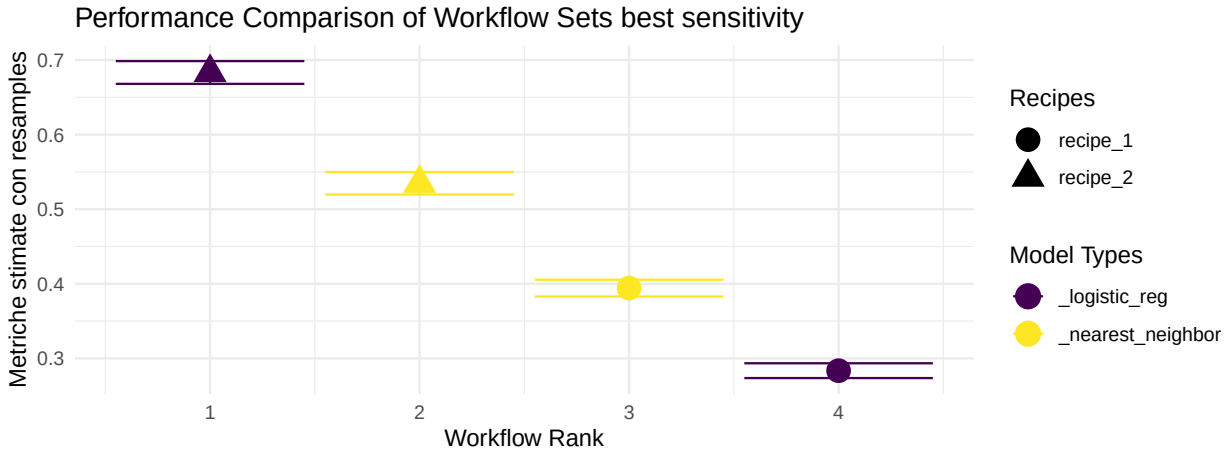

Figure 9: Capacità identificare classe minoritaria tra modelli e preprocessing- Training set

In questa valutazione infatti il modello migliore rimane la regressione logistica ma con associato un prepro-

cessing che risolve il problema dello sbilanciamento. Registra infatti un valore di sensitività pari a 0.683. In particolare se si controllano i parametri ottimizzati risultano migliori le combinazioni con il ribilanciamento al 100%.

Se infine valutiamo la ROC AUC abbiamo i seguenti risultati mostrati in figura 10.

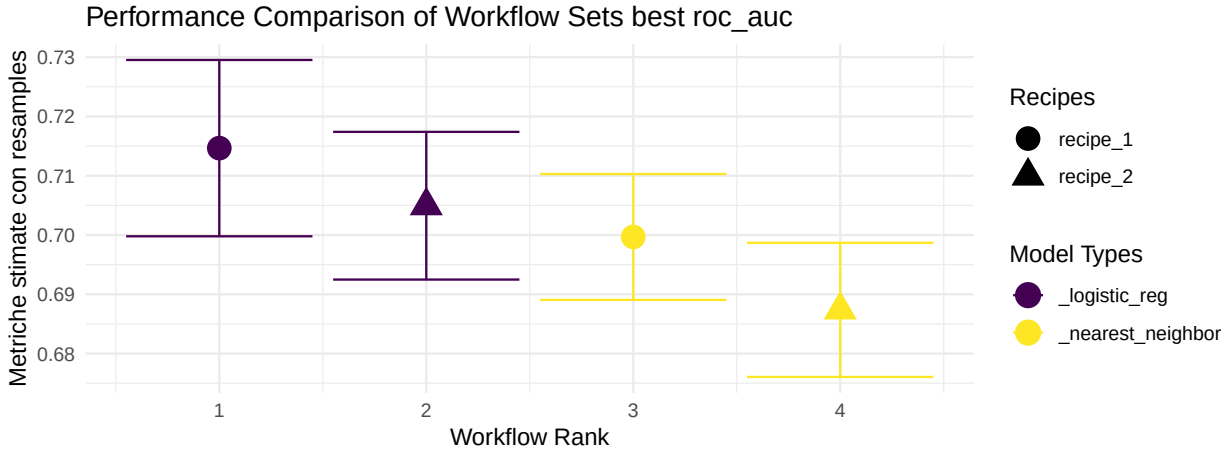

Figure 10: Efficacia al variare del cut-off tra modelli e preprocessing- Training set

La migliore area under the roc curve la registra sempre il modello logistico con pre processing di base.
Al secondo posto si posiziona il modello che era migliore a riconoscere la classe minoritaria che registra una media del 70.4% appena sotto al migliore risultato(71.4%). Per questo motivo il modello regressione logistica con upsampling della classe minoritaria viene preferito rispetto alle altre combinazioni.
In figura 11 si mostrano le performance delle 4 metriche del modello finale con i parametri ottimizzati. I risultati sono stati validati sul test set mai utilizzato da inizio analisi.

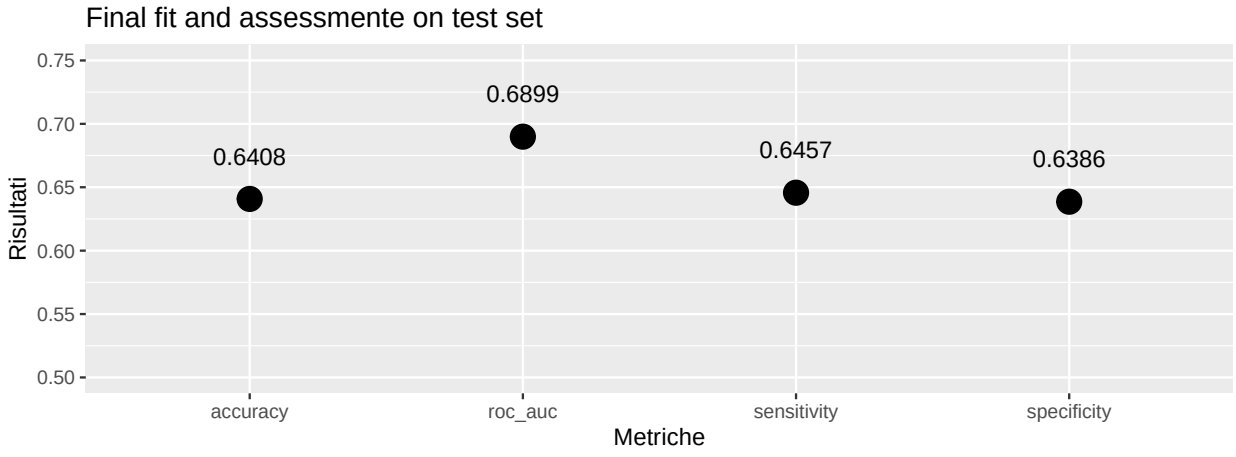

Figure 11: Metriche del modello finale scelto con parametri ottimizzati e valutato su test set

2.4 Conclusioni

Il modello di regressione logistica combinato con sovra campionamento della classe minoritaria risulta migliore al modello k nearest neighbor.
I risultati del test set sul modello finale non si discostano da quelli ottenuti in fase di training. Non si è incappati in overfitting.
Ci sono molte direzione in cui ampliare l' analisi. In fase di pre processing si potrebbe rimaneggiare le variabili sulla base di una conoscenza maggiore nel settore investing.
Oppure applicare feature reduction tramite PCA o LDA.Probabilmente una riduzione delle features favorirebbe sopratutto il metodo dei nearest neighbor che è meno performanete su dimensionalità elevate.
In fase di analisi si possono testare metodi diversi di risoluzione dello sbilanciamento di classe oppure si potrebe migliorare il modello di regressione logistica applicando opportune trasformazione dei dati.
Si potrebbero valutare anche dei modelli sulla base di società appartenenti allo stesso settore.