ANNO ACCADEMICO 2021/2022
DATA SCIENCE PER L'ECONOMIA E LE IMPRESE
DATA ANALYSIS FOR BUSINESS DECISIONS

Assignment 2

Previsione dell'EPS aziendale a 12 mesi

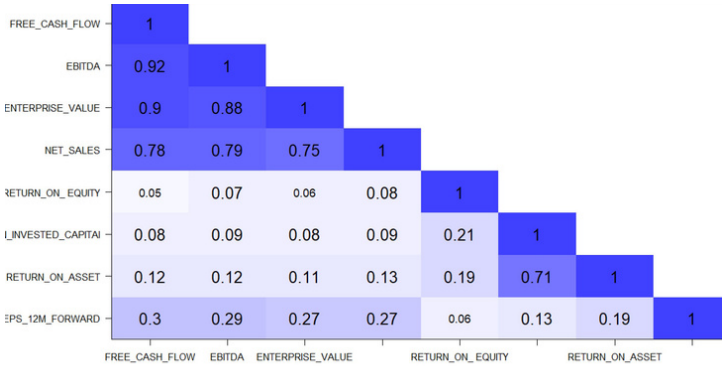
A CURA DI: RICCARDO ROCCHI, BENITO DEL PESCHIO, CHIARA ANTOGNOZZI E FEDERICA VOLPI

Obiettivo

Lo scopo della seguente analisi è quello di effettuare una previsione del valore dell'EPS a 12 mesi, sulla base di un set di variabili economiche finanziarie di 2494 diverse imprese. Oltre quest'ultime, il dataset mette in evidenza lo specifico settore di appartenenza di ciascuna impresa.
In particolare, l'analisi si propone di individuare e analizzare le variabili che incidono maggiormente ai fini di una corretta previsione.

Metodologia

Prima di procedere con la previsione dell'EPS a 12 mesi, è risultato fondamentale effettuare un'efficace analisi esplorativa del dataset, con conseguente feature selection, che ha condotto alle seguenti considerazioni:

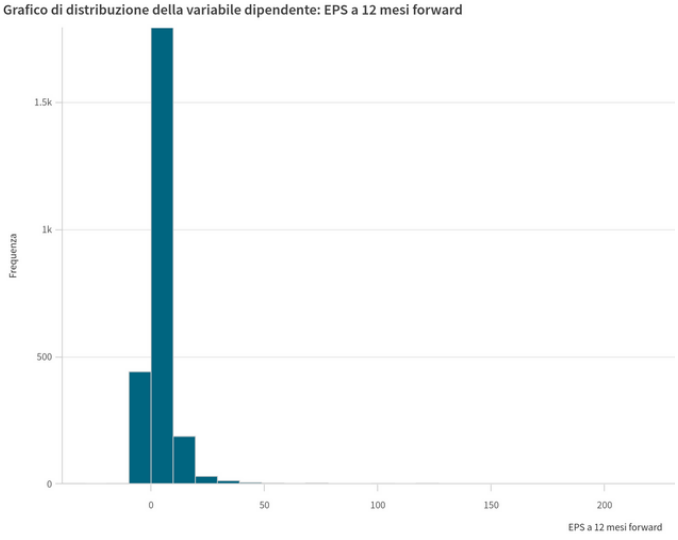


- la matrice di correlazione fornisce già da subito una possibile interpretazione delle variabili che impattano maggiormente sul valore dell'EPS (EBITDA, Free Cash Flow, Enterprise Value, Net Sales e Return on Asset);
- la regolarizzazione L1 (Lasso) ha evidenziato l'importanza di sole tre variabili per la stima dell'EPS, quali Free Cash Flow, EBITDA e Return on asset. Questo risultato si discosta leggermente da ciò inizialmente ipotizzato per due principali motivi:
  - l'assenza di Return on Invested Capital, perché solitamente usato come indice di redditività dell'impresa;
  - la presenza di un'elevata correlazione tra due variabili (EBITDA e Free Cash Flow), problematica che viene solitamente ovviata dal Lasso.

Sebbene il Lasso abbia mantenuto tre variabili, si è optato per escludere dall'analisi Free Cash Flow a causa della sua forte correlazione con l'EBITDA. Tale correlazione è motivata dal fatto che entrambe sono utilizzare per valutare la capacità di un'impresa di avere un ritorno da un business. La scelta è ricaduta sull'EBITDA in quanto indice maggiormente utilizzato per confrontare la redditività tra imprese.

Sulla base delle due variabili selezionate, sono stati individuati delle imprese che risaltano notevolmente rispetto alla maggior parte. In particolare, come si può notare dai grafici in appendice, a valori anomali corrispondono imprese da considerarsi leader nel proprio settore.

La presenza di outliers emerge anche dal qui riportato grafico di distribuzione della variabile dipendente "EPS a 12 mesi", che mostra come la maggior parte dei valori si concentri in un intervallo ristretto.



L'analisi

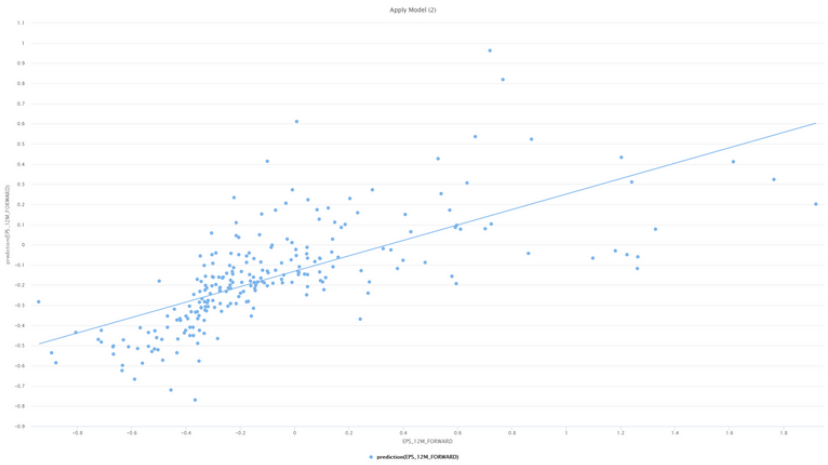
Al fine di analizzare quali variabili impattano maggiormente sul valore dell'EPS a 12 mesi, una prima tecnica utilizzata è stata quella della regressione OLS, da cui sono emersi residui eteroschedastici, come confermato dal Breusch-Pagan test. Questo problema deriva probabilmente dalla presenza di valori anomali, che peggiorano le performance della regressione.
Di conseguenza, per ovviare la problematica senza però eliminare del tutto tali valori, la scelta è ricaduta sull'utilizzo di una regressione lineare robusta, che permette di pesare differentemente le osservazioni, sulla base del comportamento più o meno regolare delle stesse. Di seguito i risultati ottenuti:

```
Residuals:
     Min       1Q   Median       3Q      Max
-5.25774 -0.13056 -0.01064  0.18987 37.89324

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.107249   0.009972 -10.755  < 2e-16 ***
RETURN_ON_ASSET  0.147750   0.010451  14.138  < 2e-16 ***
EBITDA           0.331492   0.047916   6.918 6.52e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 0.2106
Multiple R-squared:  0.4084,    Adjusted R-squared:  0.4077
Convergence in 28 IRWLS iterations
```

Il modello, oltre a restituire il livello di significatività dei coefficienti, fornisce anche un loro stima. Si può osservare che l'EBITDA è la variabile che maggiormente influisce nella stima dell'EPS, in quanto il coefficiente associatogli, oltre ad essere positivo, è il maggiore in valore assoluto. Tale risultato sembra essere coerente e realistico, considerando il dominio di applicazione. È verosimile, infatti, che L'EBITDA, essendo indice di redditività dell'impresa, impatti significativamente sul guadagno per azione.
Nonostante questa regressione lineare robusta consenta di interpretare i coefficienti, non fornisce performance out of sample soddisfacenti, restituendo un R^2 del 17.6%.

D'altro canto, altri tipi di algoritmi potrebbero risultare idonei per la previsione dell'EPS a 12 mesi, seppur non capaci di stimare coefficienti interpretabili come quelli di una combinazione lineare,
In particolare, l'SVM ha restituito buoni risultati, solo però in assenza di outliers. I record esclusi dall'analisi sono stati individuati grazie alla distanza di Cook, sulla base di una precedente regressione lineare.
La bontà della previsione out of sample è testimoniata da un R^2 del 47.5%. Tale valore è confermato dal grafico seguente, che mette in relazione valori reali e valori predetti.

Nonostante l'efficacia del modello ottenuto, si può osservare che, all'aumentare del valore dell'EPS, l'errore commesso nella stima aumenta. Ciò può essere giustificato dalla distribuzione stessa della variabile, che è particolarmente concentrata intorno a un ristretto intervallo di valori, con conseguente bassa frequenza di valori più grandi.



Appendice

Grafico a dispersione tra EPS e EBITDA: focus su valori anomali

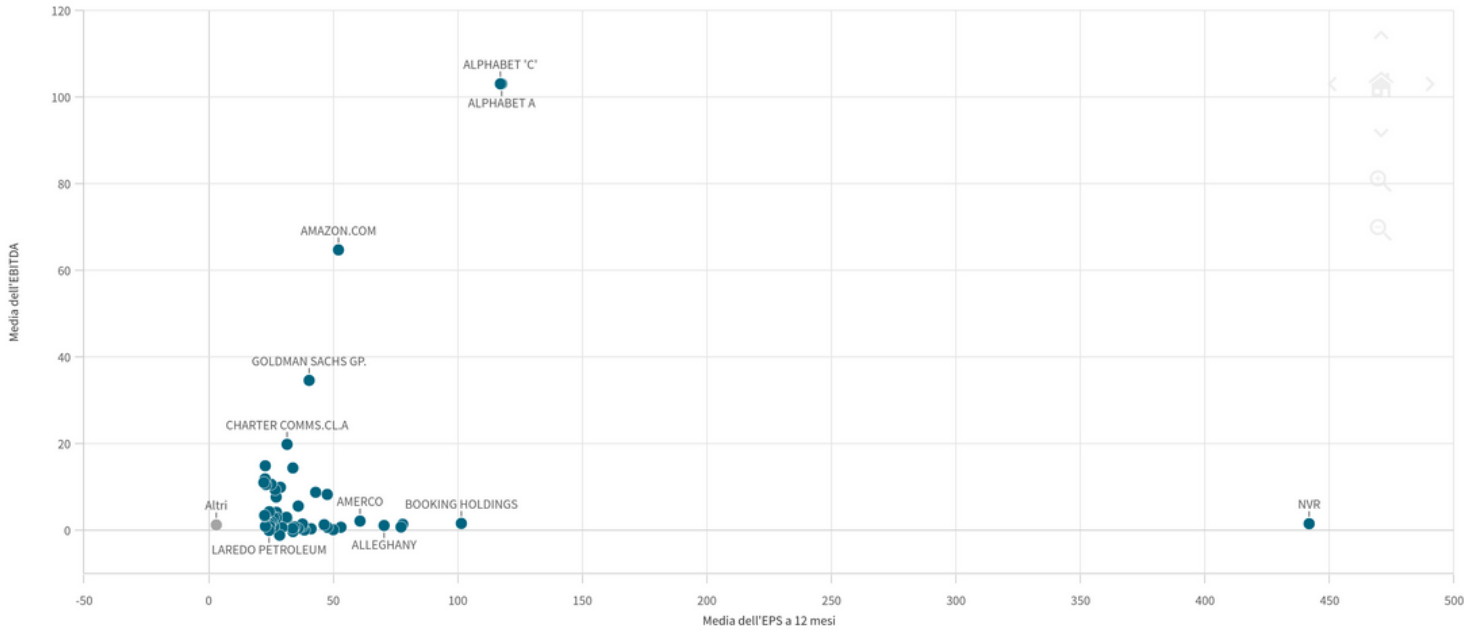


Grafico a dispersione tra EPS e Return on Asset: focus su valori anomali