PIAZZA - Report 1

Premessa

In questo report metto in evidenza i risultati e le considerazioni su quanto mi aveva chiesto nell'ultimo colloquio orale.

In particolare:

1) Aggiungere la feature *memAvg* al training dei modelli predittivi per la predizione del target *nTraces*.
2) Rafforzare il peso predittivo della feature *sol(keuro)* sfruttando le feature *Load* e *PV*.

Prima di parlare dei risultati riguardanti questi due punti, può essere di aiuto alla comprensione una considerazione riguardo la correlation Matrix tra le features, che ho provveduto a graficare.

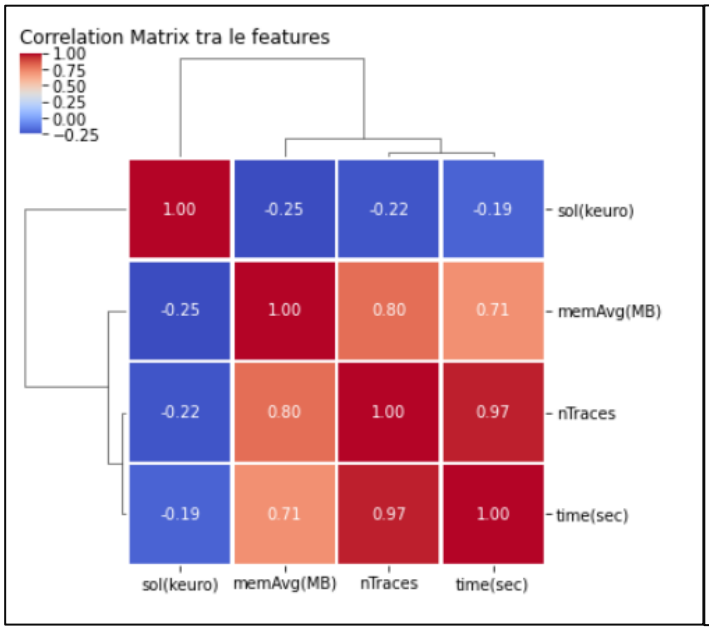


Fig. 1

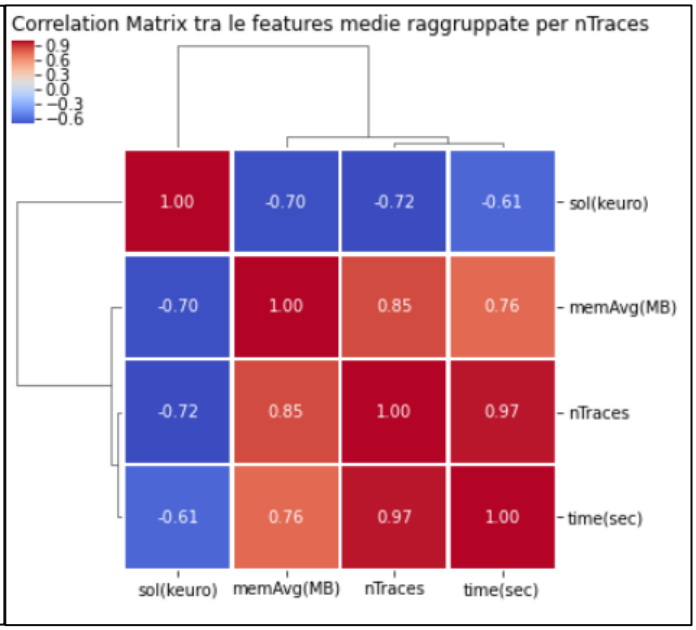


Fig. 2

Dalla prima correlation Matrix (Figura 1) possiamo notare che:

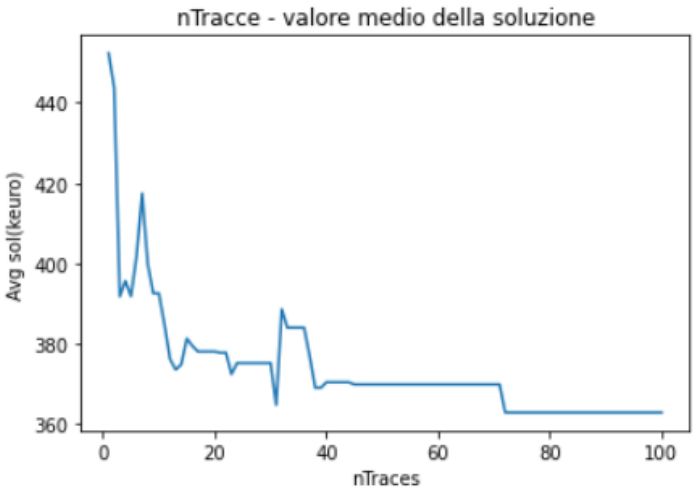
- nTraces è fortemente correlato in primo luogo con *time* e poi con *memAvg*.
- sol(keuro) è scarsamente correlata con tutte le features considerate.

Diverso è il discorso se consideriamo i valori medi delle features raggruppate per numero di tracce. In figura 2 infatti notiamo che:

- nTraces è ancora fortemente correlato in primo luogo con *time* e poi con *memAvg*, ma ora ha una discreta correlazione inversa anche con *sol*.
- aumenta la correlazione di *sol* con le altre features considerate.

Queste osservazioni giustificano la puntualizzazione che mi aveva fatto: "Come mai osserviamo una correlazione nell'andamento del grafico nTracce-soluzione media, e poi i risultati ottenuti dai modelli predittivi allenati con sol(keuro) sono scarsi?"

Il grafico in questione era il seguente:

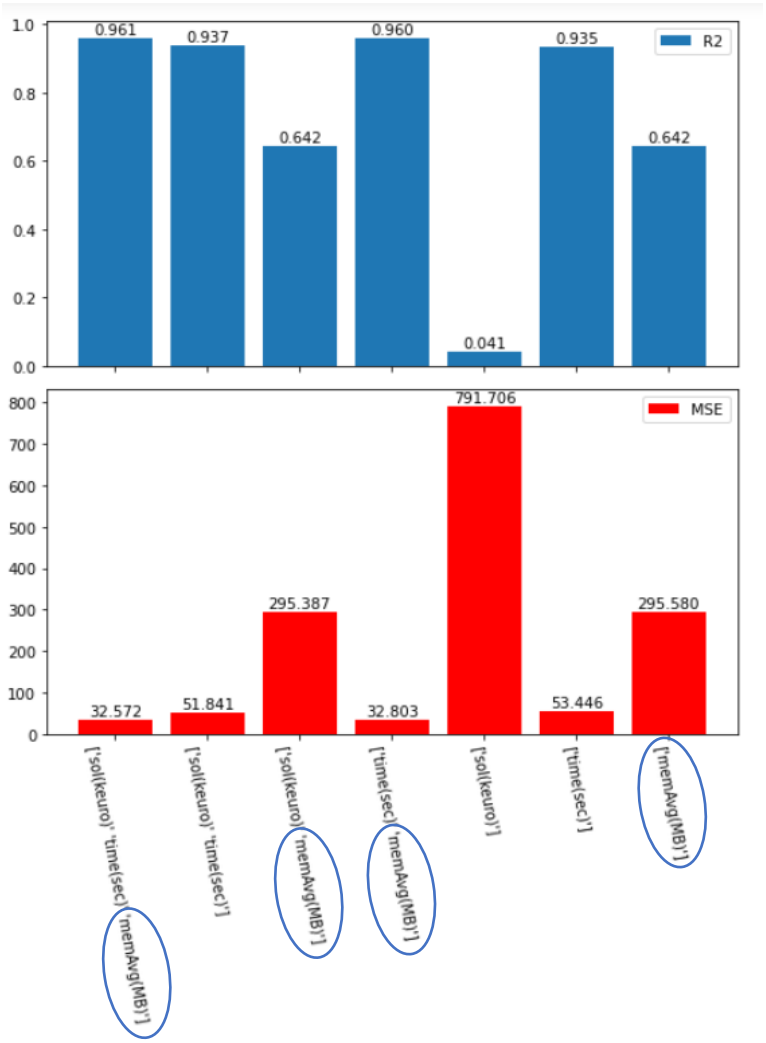


L'andamento decrescente che vediamo, infatti, è relazione tra il valore **medio** delle soluzioni, raggruppate per numero di tracce, e nTraces. Tuttavia, per allenare i modelli predittivi io ho considerato i valori di soluzione singoli che, come visto in Figura 1, sono scarsamente correlati con il target da predire (nTraces). Non ho considerato i valori medi.

Fatta questa premessa passo a commentare le considerazioni principali citate in apertura.

1) Aggiunta della feature *memAvg* al training dei modelli predittivi per la predizione del target *nTraces*.

Performance ottenute con la **Linear Regression:**



COMMENTO

I risultati ottenuti rispecchiano le considerazioni derivate dalle *correlation Matrix*.

Le feature maggiormente correlate con nTraces sono *time(sec)* e *memAvg(MB)*. Non ci sorprende dunque il fatto che **i risultati di predizione migliori sono stati ottenuti quando il modello viene allenato considerandole entrambe**. Al secondo posto nella classifica delle performance troviamo i risultati di predizione ottenuti considerando *time(sec)* ed al terzo quelli ottenuti considerando *memAvg(MB)*.

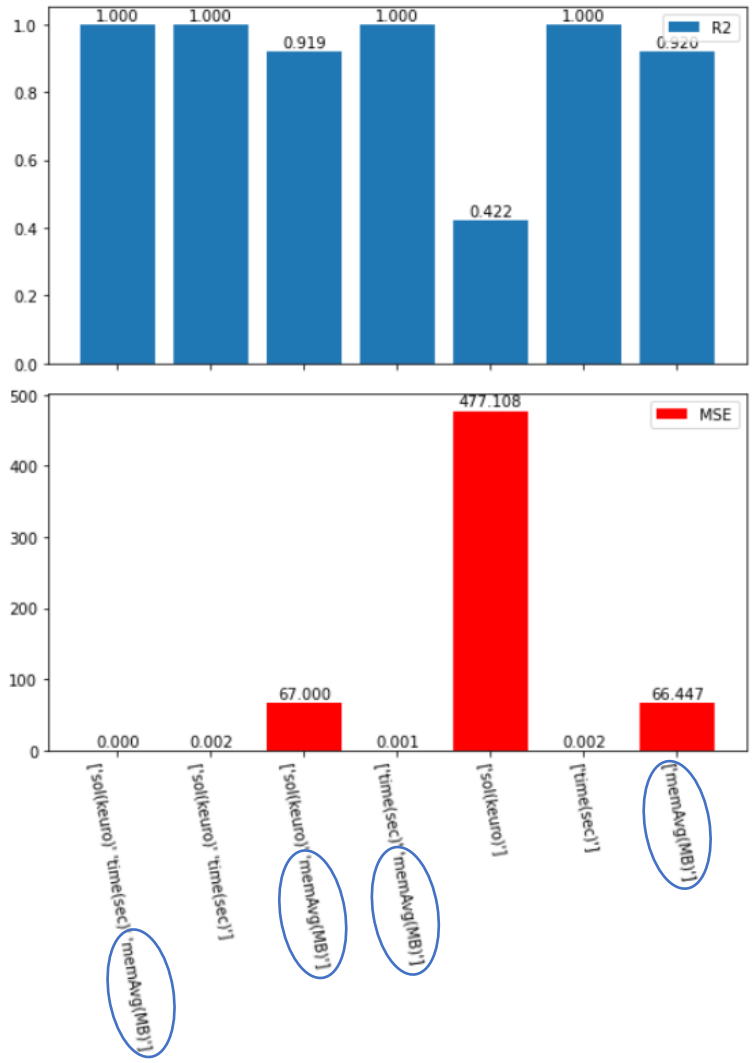
In questa classifica non è stata volutamente citata *sol(keuro)* in quanto possiamo notare che la sua presenza/assenza tra le feature su cui allenare il modello è pressoché irrilevante in termini di performance.

Esempio: La R2 sul modello di multiple regression in cui gli attributi usati per allenare il modello sono memAvg(MB), time(sec) e sol(keuro) è la migliore ottenuta (96.1%), ed è pressoché identica alla R2 ottenuta senza sol(keuro) (96%). Anche l'MSE senza sol(keuro) rimane quasi invariato (peggiora di 0.3).

Questo comportamento è giustificato dal fatto che **sol(keuro) è scarsamente correlata con nTraces** e quindi **non è molto utile alla predizione del target**.

La R2 sul modello di linear regression in cui considero solo sol(keuro) infatti è parecchio scarsa (circa 4%).

Performance ottenute con i **Regression Tree:**



COMMENTO

- I risultati ottenuti con i modelli di Regression Tree (maxDepth=10) sono migliori di quelli ottenuti dalla Linear Regression, in quanto le feature non presentano tra loro delle relazioni lineari.
- Parecchi modelli qui allenati hanno un R2 prossimo a 1 e un MSE prossimo a 0.
- Sebbene la R2 della sol(keuro) sia migliorata (ora è al 42%) vale ancora la stessa osservazione fatta per la linear Regression: **la sua presenza/assenza non inficia significativamente sulle performance predittive ottenute con le altre feature.**

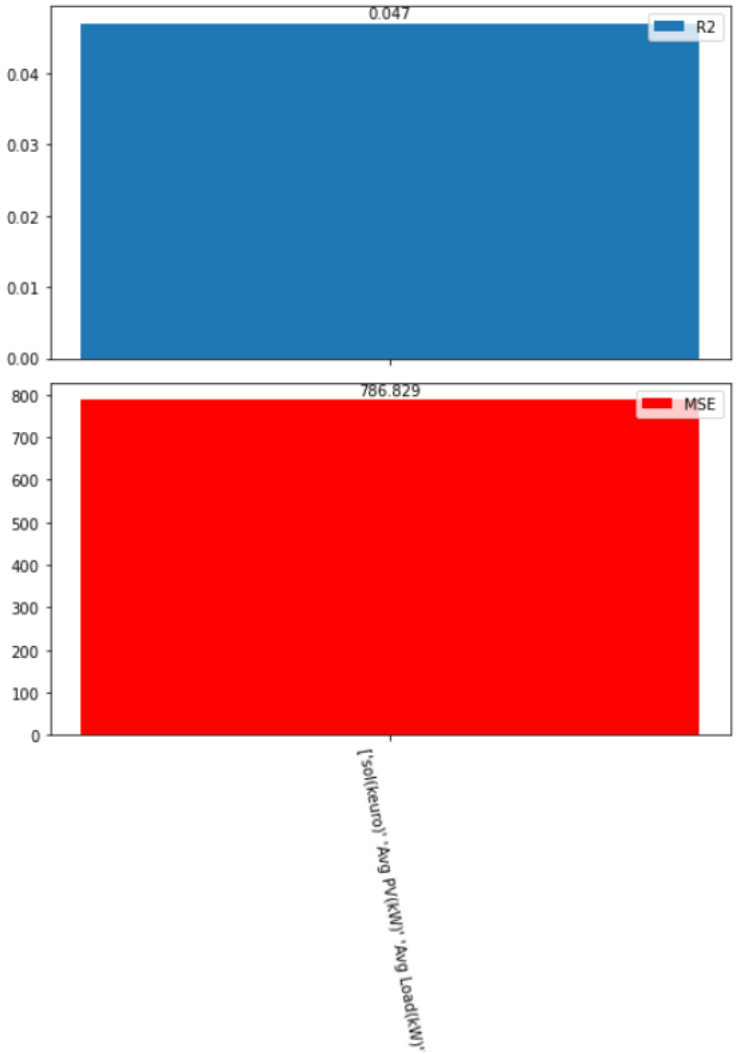
2) Rafforzare il peso della feature *sol(keuro)* sfruttando le feature *Load* e *PV*.

Motivazioni di questo test: Come evinto dalle considerazioni precedenti, per predirre in modo accurato il numero di Tracce le feature più convenienti da considerare sono *time* e *memAvg* rispetto alla feature *sol*.
Lo scenario di utilità reale che si prospetta quindi è:
"*abbiamo un certo tempo e/o una certa quantità di memoria a disposizione e dobbiamo far eseguire l'algoritmo di fixing, quante tracce possiamo fargli prendere in considerazione?*"
Tuttavia, nella realtà si potrebbero avere esigenze anche diverse.
Ad esempio: "*voglio ottenere una certa qualità di soluzione. Quante tracce devo far prendere in considerazione all'algoritmo?*"

Visto quanto realizzato fino ad ora, se si considera un modello predittivo che considera solo la feature soluzione, non si ottengono predizioni su *nTraces* molto accurate.
Lo scopo di questo paragrafo è quello di considerare insieme alla sol(keuro) anche le feature PV(kW) e load(kW), da cui essa è fortemente dipendente, ai fini di sviluppare un modello predittivo con target nTraces che sia più accurato.

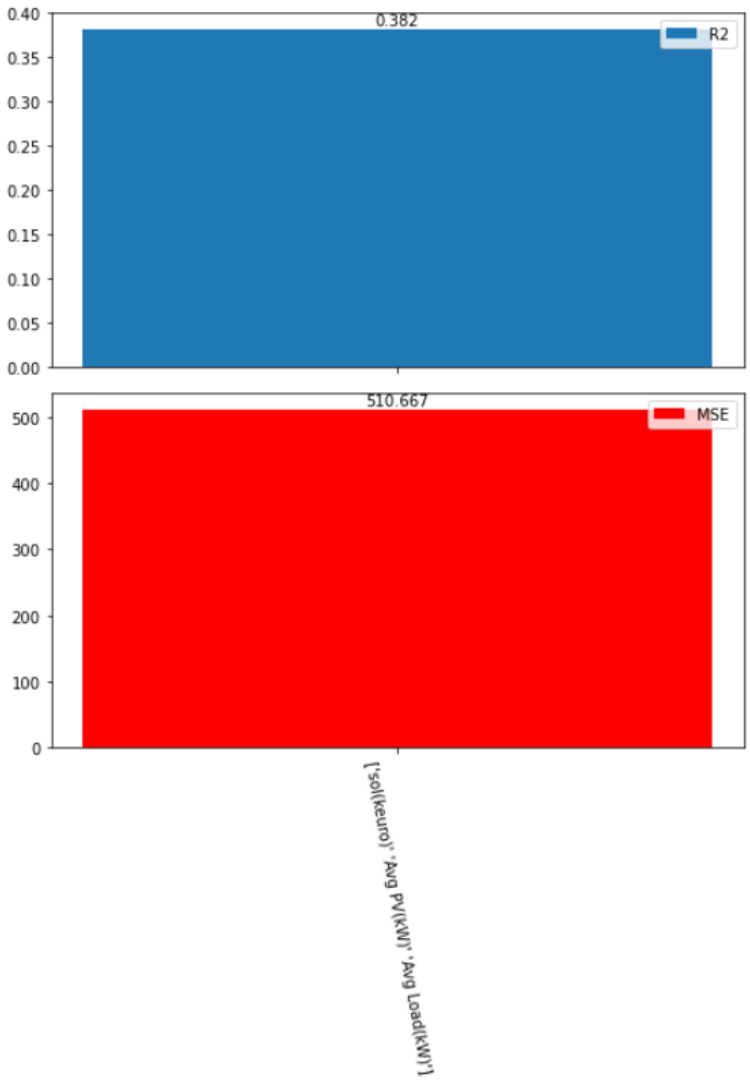
Dal momento che PV e Load sono dei vettori di 96 elementi per ogni istanza, mentre sol(keuro) è un unico valore ottenuto come somma delle 96 soluzioni sui 96 stage, ho considerato, per ogni soluzione la **media dei 96 valori di PV e la media dei 96 valori di Load.**

Performance ottenute con la **Linear regression:**



COMMENTO: Le metriche ottenute da questo modello di regressione lineare in cui sono state utilizzate le feature *soluzione*, *Load media* e *PV media* non sono significativamente migliori del corrispondente modello in cui era stata considerata solo la *soluzione*.
La R2 qui ottenuta si può considerare analoga alla R2 ottenuta precedentemente (migliora solo di 0.006%).
Discorso analogo per MSE che migliora di soli 5 punti passando da 791 a 786.

Performance ottenute con i **Regression Tree:**



COMMENTO: Le metriche ottenute da questo modello di regression tree in cui sono state utilizzate le feature *soluzione*, *Load media* e *PV media* sono addirittura peggiori del corrispondente modello in cui era stata considerata solo la *soluzione*.
La R2 qui ottenuta è peggiorata di 4 punti percentuali (è scesa da 42% a 38%).
MSE è peggiorato di una trentina di unità passando da 477 a 510.

Conclusioni

Le feature Load(kW) e PV(kW) non possono essere utilizzate per predirre nTraces in quanto quest'ultimo è totalmente indipendente da esse.
Infatti è sufficiente osservare che gli stessi valori di Load e PV si ripetono periodicamente per ogni valore di nTraces, di conseguenza non possono rappresentare un carattere utile a predirre il target.

Esempio: Dati *x*, media aritmetica di 96 valori di PV, e y, media aritmetica di 96 valori di Load, sappiamo predirre nTraces? No, in quanto ritrovando lo stesso x e la stessa y in tutti i valori di nTraces, l'entropia è massima (i valori assumibili da nTraces sono tutti equiprobabili).