Predizione della progressione del diabete con il Machine Learning

Federico Trotta

PROGETTO MACHINE LEARNING

Descrizione del progetto

Le specifiche del progetto erano quelle di analizzare il dataset "diabetes" della libreria scikit-learn e di scegliere un modello (qualsiasi) di Machine Learning per la predizione della progressione della malattia, con l'unico vincolo di scegliere un modello con almeno un iperparametro da validare

Non mi sono accontentato, e ne è nato uno studio di diversi modelli di Machine Learning per prendere confidenza coi risultati delle metriche; ho, quindi, imparato a capire come valutare i risultati numerici delle varie metriche utilizzate per la validazione dei modelli

Il modello di Regressione Lineare

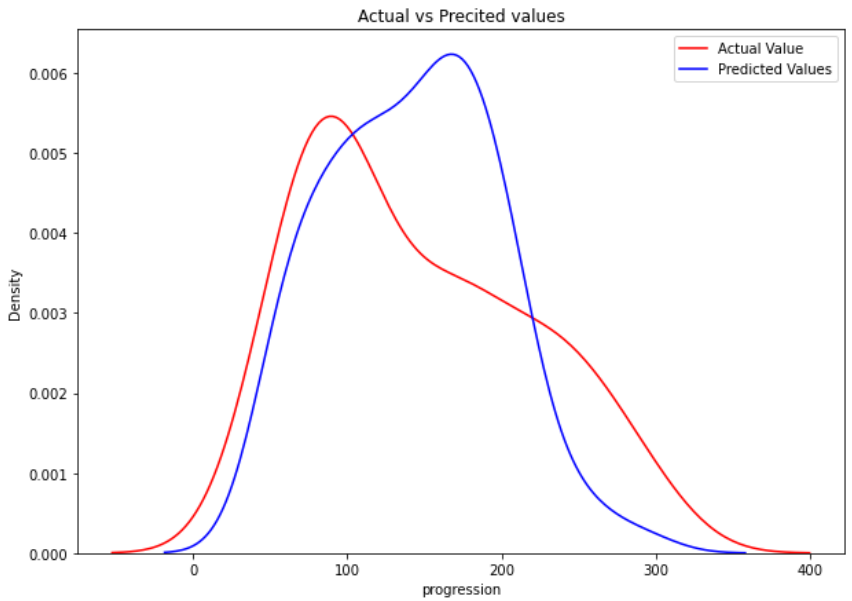
Sono partito dal modello di regressione lineare, studiando il MAE , il RMSE ed il coefficiente di determinazione come metriche. Ho poi usato il grafico KDE per la validazione finale. I risultati non erano troppo soddisfacenti, quindi sono passato al modello regolarizzato.

```
#model metrics
print(f'The mean absolute error is:{metrics.mean_absolute_error(y_test, y_test_pred): .2f}')

print(f'The root mean squared error is:{np.sqrt(metrics.mean_squared_error(y_test, y_test_pred)): .2f}')
```

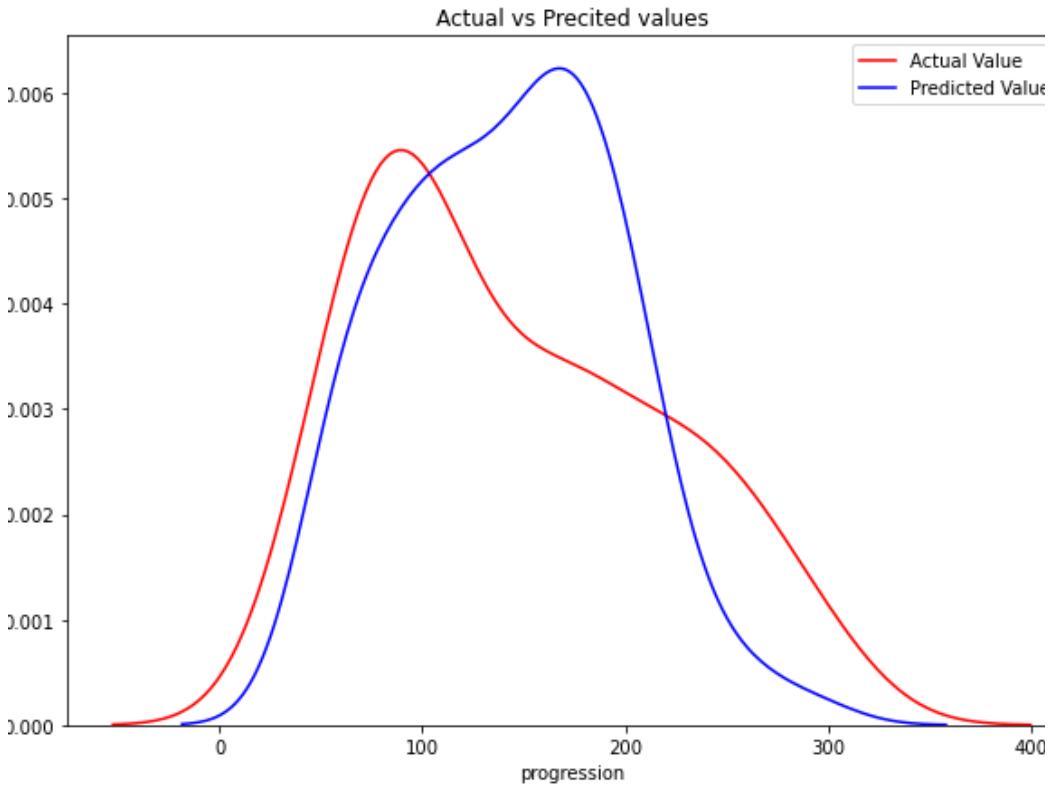
```
The mean absolute error is: 42.79
The root mean squared error is: 53.85
```

These are high values, which continue to make me feel discouraged about the use of this model ... but I want to make a couple of visualizations.



Il modello di regressione lineare regolarizzato di tipo Lasso

Ho provato ad applicare la regolarizzazione di tipo Lasso, ma i risultati ottenuti sono stati pressoché identici a quelli del modello non regolarizzato

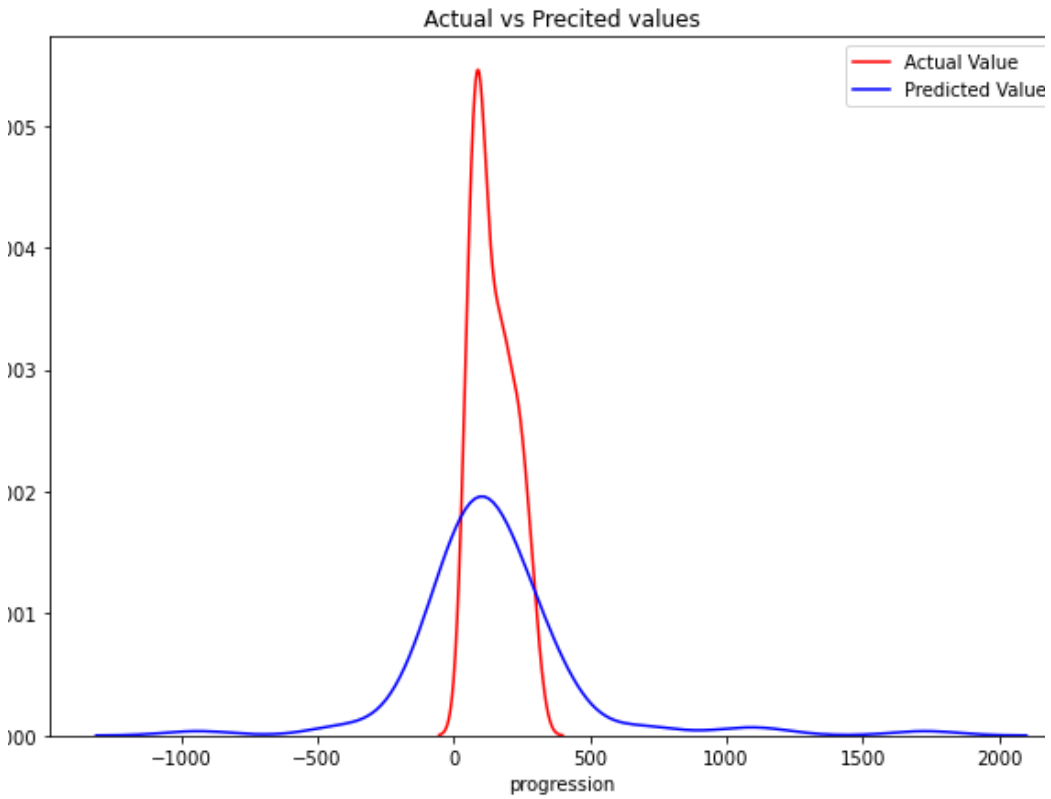


Il modello di regressione polinomiale

Successivamente, ho tentato il modello di regressione polinomiale di terzo grado, ma ho ottenuto overfitting

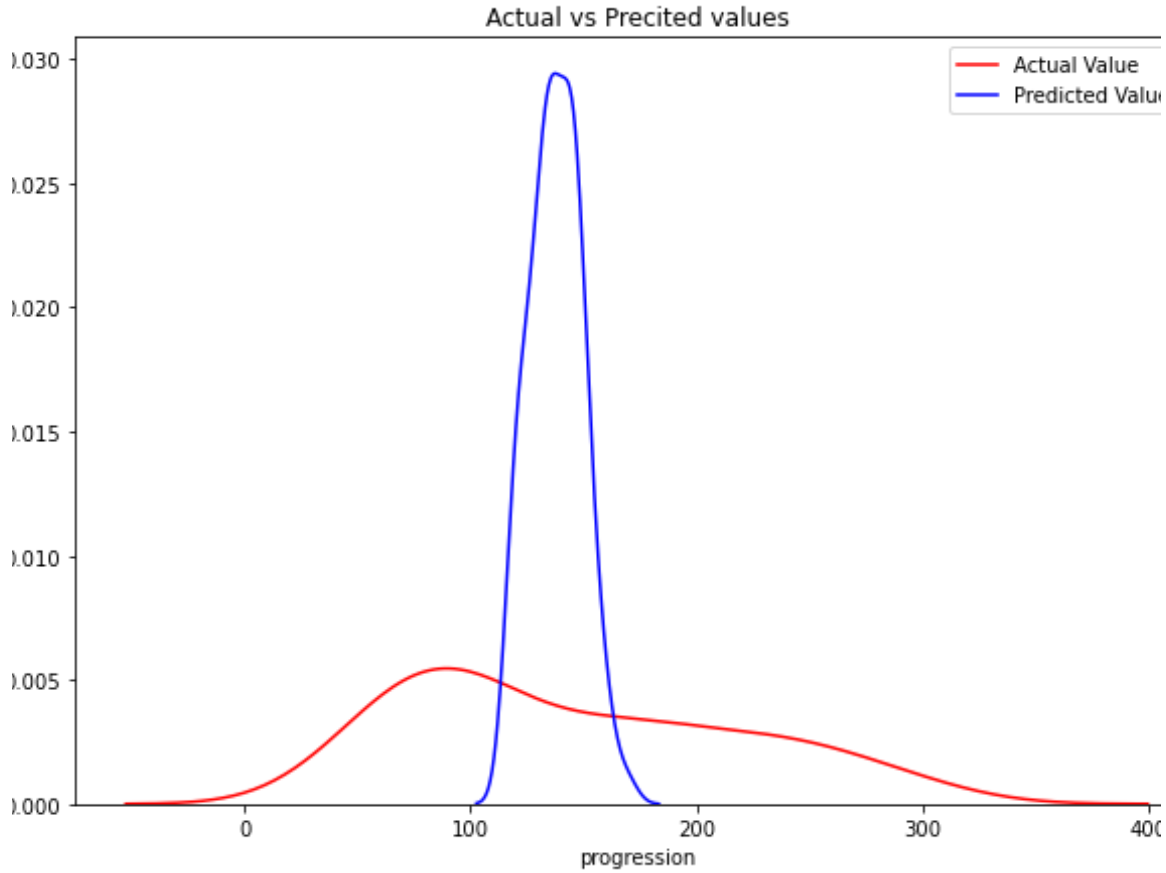
```
#R^2
print(f'Coeff. of determination on train set:{poly_reg.score(X_train3, y_train3): .2f}') #train set
print(f'Coeff. of determination on test set:{poly_reg.score(X_test3, y_test3): .2f}') #test set
```

```
Coeff. of determination on train set: 0.88
Coeff. of determination on test set:-17.42
```



Il modello Support Vector Regression (SVR)

Infine, ho tentato il modello SVR che, però, è risultato fin da subito il meno adatto, in quanto ho ottenuto dei valori del coefficiente di determinazione praticamente nulli. Il grafico KDE ha poi confermato questa idea iniziale.



Conclusioni

Questo progetto è stato un modo per prendere confidenza con le varie metriche tipiche dei problemi di regressione.

Ho imparato a validare gli iperparametri ed a confrontare vari modelli di ML sulla base delle metriche.

Oltretutto, il modello di regressione lineare semplice risulta non essere troppo ottimale, ma alla fine è risultato il migliore di tutti quelli che ho provato: questo mi ha insegnato ad imparare a gestire i miei bias nello studio di un problema di ML.

Per vedere il progetto completo

Ho svolto una analisi più completa, scrivendo due articoli per "Towards Data Science". Qui la prima parte e qui la seconda.

Il progetto completo si trova sul mio profilo GitHub qui