Machine learning

Davide Abete, Fabrizio Cominetti e Agazzi Ruben

January 8, 2022

Contents

Il progetto consiste nell'analisi di vari modelli di regressione applicati ad un dataset contenente le informazioni ambientali riferite alla temperatura globale a partire dal Gennaio 1850 al Novembre 2015.

1 Introduzione

Il cambiamento climatico è reale. La sfida è avvincente. E più a lungo aspettiamo, più difficile sarà risolvere il problema.[4]

L'ultimo secolo si è reso protagonista di un aumento graduale e preoccupante della temperatura globale, ed oggi questo tema ha assunto rilevanza cruciale in numerosi dibattiti politici relativi al futuro del pianeta.Scopo di questo progetto è di realizzare modelli di machine learning con il fine di prevedere le future temperature medie globali. Precisamente, saranno utilizzati modelli di regressione per prevedere la temperatura del mese successivo a quello preso in considerazione. (Lo scopo di questo lavoro consiste nel realizzare e successivamente analizzare vari modelli di regressione con il fine di prevedere la temperatura media globale del mese successivo, sulla base dei dati rilevati durante la mensilità precedente. Come specificato in apertura di sezione, la scelta di questo progetto è stata influenzata dal crescente interesse verso le tematiche riguardanti il cambiamento climatico, sia da

un punto di vista personale che, appunto, globale. I dati utilizzati per realizzare questo progetto sono scaricabili dalla piattaforma Kaggle al seguente indirizzo () e contengono diverse misure di temperature medie, rilevate ogni mese a partire dal Gennaio 1750 al Novembre 2015.

2 Dataset

Il dataset utilizzato è intitolato "Climate Change: Earth Surface Temperature Data"[3].Come precisato poco sopra, il dataset è stato ottenuto tramite la piattaforma Kaggle, sulla quale è stata resa disponibile una versione ripulita del dataset fornito dalla "Berkeley Earth", un'organizzazione non-profit indipendente specializzatta in temi di 'environmental data science'. L'organizzazione Berkeley Earth definisce il problema del global warming come 'la sfida definitiva del nostro tempo' e pone l'accento sulla necessità di ottenere con urgenza dati di qualità e informazioni scientifiche di valore sul tema. Il progetto è stato svolto principalmente tramite l'utilizzo della piattaforma Knime e, in particolari casi, del linguaggio di programmazione Python.

2.1 Descrizione del dataset

Il dataset è composto da 3192 record e dai seguenti 9 attributi:

- dt: Data di rilevamento dei dati
- landAverageTemperature: temperatura media globale del terreno espressa in gradi Celsius.
- LandAverageTemperatureUncertainty: Intervallo di confidenza al 95% intorno alla media della temperatura globale
- LandMaxTemperature: media della temperatura massima globale del terreno espressa in gradi Celsius
- LandMaxTemperatureUncertainty: Intervallo di confidenza al 95% intorno alla media della temperatura massima globale
- LandMinTemperature: media della temperatura minima globale del terreno espressa in gradi Celsius
- LandMinTemperatureUncertainty: Intervallo di confidenza al 95% intorno alla media della temperatura minima globale
- LandAndOceanAverageTemperature: Temperatura media globale del terrestre e oceanica espressa in gradi Celsius.
- LandAndOceanAverageTemperatureUncertainty: Intervallo di confidenza al 95% intorno alla media della temperatura media terrestre e oceanica globale.

2.2 Esplorazione dei dati

Per la fase di esplorazione dei dati è stato utilizzato principalmente il nodo Statistics, presente all'interno della piattaforma Knime[2], attraverso il quale sono stati rilevati alcuni indici di posizione, variabilità e il numero di missing values, dei quali ne elencheremo i principali di seguito:

Colonna	Media	STD	Missing V.
Col. 1	8.3747	4.3813	12
Col. 2	0.9385	1.0964	12
Col. 3	14.3506	4.3096	1200
Col. 4	0.4798	0.5832	1200
Col. 5	2.7436	4.1558	1200
Col. 6	0.4318	0.4458	1200
Col. 7	15.2126	1.2741	1200
Col. 8	0.1285	0.0736	1200

La fase di esplorazione ha permesso di identificare la variabile target che ci interessa prevedere, ovvero *landAverageTemperature* del mese successivo, e le colonne restanti che fungono da attributi esplicativi. La variabile target selezionata è di tipo numerico, continua e assume valori fra $[-2.080, 19.021]$ con 3 cifre decimali.

3 Pre-processing

Durante la fase di data exploration si è potuto osservare che molte colonne presentavano un

numero di valori mancanti pari a 1200. Ciò è dovuto al fatto che per le osservazioni effettuate prima del Gennaio 1850 non sono stati rilevati tali dati di interesse.

3.1 Missing values

Per risolvere questa problematica si è deciso di rimuovere i record che presentano missing values, eliminando quindi 1200 righe e riducendo l'intervallo di tempo dei valori che, a questo punto, partono dal Gennaio 1850 invece che dal Gennaio 1750. Per eliminare le righe che presentano missing values è stato utilizzato il nodo di Knime chiamato *«Missing Values»*

3.2 Data augmentation

Per essere in grado di allenare i vari regressori è necessario il dato relativo alla temperatura media terrestre del mese successivo. Per fare questo abbiamo ordinato in modo decrescente i dati in base alla colonna contenente la data utilizzando il nodo *«Sorter»* e in seguito abbiamo utilizzato il nodo *«Lag Column»* per creare una nuova colonna contenente la temperatura media del mese successivo.

3.3 Selezione delle variabili

Per selezionare le variabili da utilizzare per l'apprendimento dei regressori è stato utilizzato un filtro di correlazione in modo da eliminare attributi ridondanti. Il valore soglia di correlazione scelto è di 0.9. Alla fine del processo di feature selection sono state tenute 6 colonne su 10. Per effettuare questa operazione sono stati utilizzati i nodi *«Linear Correlation»* e *«Correlation Filter»*. Alla fine del processo di *Feature Selection* le variabili utilizzate per allenare i vari regressori sono:

- LandAverageTemperature
- LandAverageTemperatureUncertainity
- LandMaxTemperatureUncertainity
- LandMinTemperatureUncertainity

4 Modelli

Per effettuare la previsione della temperatura media terrestre globale del mese successivo, sono stati usati vari modelli di regressione:

- Regressione Lineare
- Regressione Polinomiale
- XgBoost
- Tree Ensemble
- Random Forest

4.1 Regressione Lineare

Il modello di regressione lineare consiste in una funzione lineare, che prende in ingresso le variabili esplicative, la cui risposta è il valore previsto della variabile target[1]. Nel nostro caso utilizziamo le variabili esplicative selezionate tramite feature selection per predire la variabile *LandAverageTemperature*

4.2 Regressione Polinomiale

Il modello di regressione polinomiale, a differenza di quello lineare, prevede il valore della variabile target tramite una funzione polinomiale, in cui si può specificare il grado della funzione[1]. Nel nostro caso è stato scelta un polinomio di quarto grado.

4.3 XgBoost

XgBoost è un implementazione di alberi di decisione con gradient boosting. Gli alberi di decisione sono aggiunti uno alla volta all'insieme di alberi, e configurati in modo da correggere gli errori di previsione degli alberi precedenti[5]. I vantaggi di XgBoost consistono nel fatto che l'apprendimento del modello è parallelizzabile, e quindi molto veloce; gestisce autonomamente eventuali missing values e inoltre previene l'overfitting del modello tramite *Regularization*

4.4 Tree Ensemble

Il modello di regressione Tree Ensemble allena un insieme di alberi di regressione. Tipi-

camente ogni albero è allenato utilizzando un diverso set di righe e/o colonne del dataset. Il valore di un nodo foglia di un albero di decisione consiste nella media della variabile target dei record del dataset contenuti nel percorso del nodo foglia. Detto questo l'output di un modello di regressione Tree Ensemble è la media delle previsioni ottenute dai vari alberi di regressione[2].

4.5 Random Forest

Il modello di regression Random Forest è un tipo particolare di modello Tree Ensemble. In particolare il Random Forest, a differenza del modello Tree Ensemble, seleziona le variabili esplicative, usate per l'allenamto dei singoli alberi, in modo casuale[2].

5 Training

Per la fase di training dei modelli di regressione è stato utilizzato l'approccio della *Cross Validation*.

5.1 Cross Validation

La Cross Validation consiste nel dividere i dati di training in *K* "Fogli". Per ogni foglio $k \in \{1, ..., K\}$, eseguiamo l'allenamento del modello con utilizzando $K - 1$ fogli, ed eseguiamo il test sul foglio rimanente; inoltre viene eseguita questa operazione a rotazione, con numero di ripetizioni uguale al numero di fogli, in modo da eseguire il training e il testing utilizzando tutti i sottoinsiemi. Infine si calcola l'errore facendo una media fra gli errori riscontrati durante ogni ciclo di training-testing. L'approccio della cross validation è molto utile in quanto permette di allenare e testare i vari regressori con tutti i dati presenti nel dataset. Nel nostro caso abbiamo utilizzato 10 fogli, i vari record sono scelti tramite random sampling. Per l'impletazione della cross validation è stato creato per ogni modello un meta nodo, con all'interno un nodo x-partitioner, il quale effettua la divisione in "fogli", e un nodo x-aggregator per aggregare i vari risultati; inoltre, siccome il nodo x-aggregator non fornisce i valori intermedi di R^2 relativo ad ogni ciclo di cross validation, abbiamo inserito un nodo di tipo *Python Script* per fornire in output queste informazioni mancanti.

6 Valutazione

Per la fase di valutazione dei vari modelli sono state prese in considerazione diverse metriche per ogni modello e, inoltre, è stata effettuata un analisi dei residui dove possibile.

6.1 Analisi r-quadro

L'R^2, anche detto *coefficiente di determinazione*, rappresenta in percentuale la proporzione della variazione della variabile dipendente predetta utilizzando le variabili indipendenti [6].

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Per tutti i modelli è stata eseguita una comparazione delle varie misure di R^2 intermedie ad ogni passo della cross validation, ottenendo i seguenti risultati:

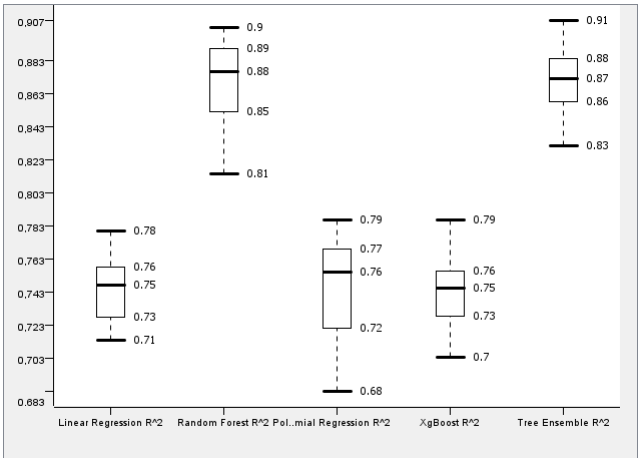

Figure 1: *Box plot relativo ai vari valori di R^2 di ogni ciclo della cross validation di ogni modello.*

Come si può osservare dal box plot i valori migliori di R^2 sono stati ottenuti dai modelli

Random Forest e *Tree Ensemble*, i quali hanno valori di R^2 compresi fra $[0,81, 0,91]$. Il fatto che questi due modelli abbiano valori simili era prevedibile, in quanto il modello *Random Forest* è una variante particolare del modello *Tree Ensemble*. Il valore peggiore di R^2 appartiene al modello di *Regressione Polinomiale*, il quale, oltre ad avere una valore mediano pari a $0,76$, ha una grande varianza rispetto ai valori di R^2 dei vari cicli di cross validation.

6.2 Analisi Mean Squared Error

Il Mean Squared Error midura la meedia dedi quadrati degli errori presenti nelle previsioni del modello[7].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Per tutti i modelli presi in considerazione è stato calcolato il malore di *Mean Squared Error* ad ogni ciclo di cross validation. I risultati ottenuti sono i seguenti:

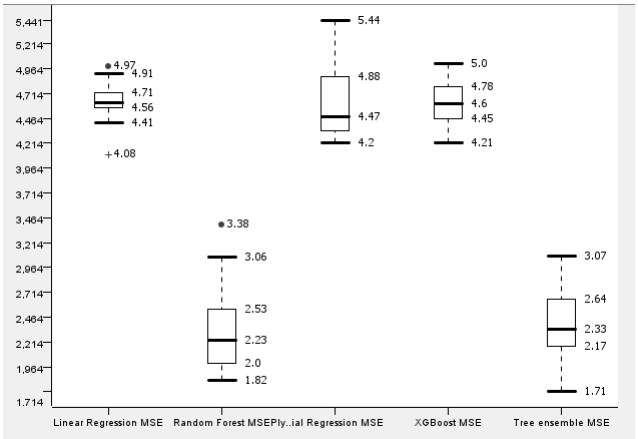

Figure 2: *Box plot relativo ai vari valori di MSE di ogni ciclo della cross validation di ogni modello.*

Come si può notare dal grafico i modelli con *MSE* migliore, quindi più basso, sono il *Tree Ensemble* e *Random Forest*, i quali hanno valori che variano nell'intervallo $[1,71, 3,38]$. I modelli *Linear Regression, Polinomial Regression e XgBoost* sono quelli con *MSE* peggiore, con valori che variano nell'intervallo $[4,08, 5,44]$.

6.3 Analisi dei residui

Per l'analisi della qualità di modelli di tipo *Linear Regression e Polinomial Regression*, un passo fondamentale consiste nell'eseguire l'analisi dei residui. Questa analisi consiste nell'osservare il tipo di distribuzione assunto dai residui ottenuti nella previsione della variabile target: devono avere una distribuzione di tipo normale. Per verificare la loro distribuzione è possibile effettuare una prima analisi visiva tramite l'utilizzo di istogrammi, e in seguito utilizzare test d'ipotesi, come ad esempio il test Shapiro-Wilk, per verificare la distribuzione.

6.3.1 Regressione Lineare

In seguito ad un analisi visiva dell'istogramma relativo ai residui del modello di Regressione Lineare è già possibile notare che i residui non sono distribuiti normalmente:

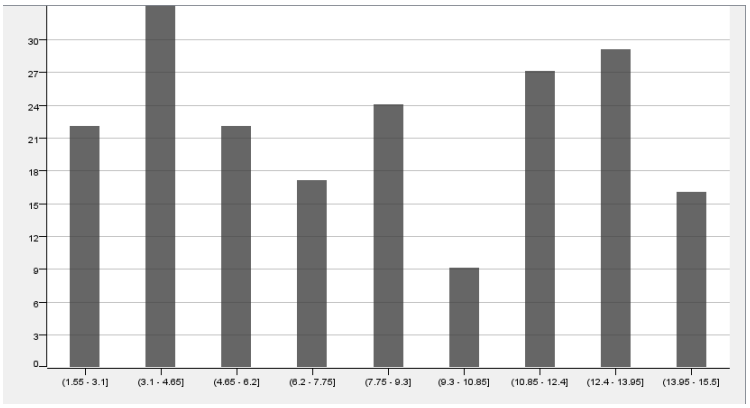

Figure 3: *Istogramma relativo ai residui del modello di Regressione Lineare, si può osservare visivamente la non Normalità della distribuzione dei residui.*

Per confermare l'ipotesi visiva è stato eseguito il test d'ipotesi Shapiro-Wilk, il quale ha rigettato l'ipotesi che i campioni relativi ai residui utilizzati dal test siano distribuiti seguendo una distribuzione normale.

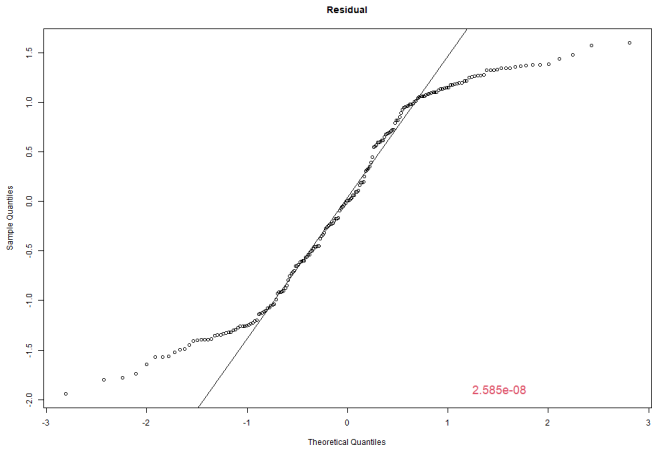

Figure 4: *Q Q Plot relativo ai residui del modello di Regressione Lineare*

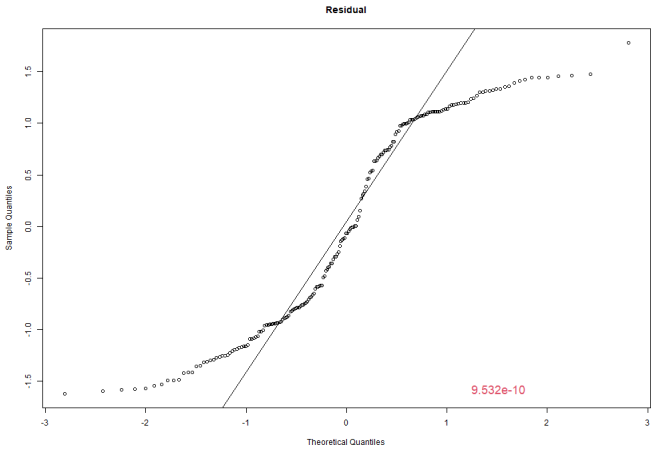

Figure 6: *Q Q Plot relativo ai residui del modello di Regressione Polinomiale.*

6.3.2 Regressione Polinomiale

In seguito ad un analisi visiva dell'istogramma dei ai residui del modello di Regressione Polinomiale si può già stabilire che i residui del modello non sono distribuiti normalmente:

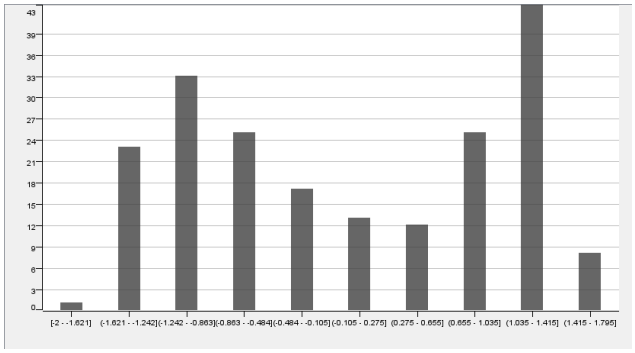

Figure 5: *Istogramma relativo ai residui del modello di Regressione Polinomiale.*

Per confermare l'ipotesi visiva è stato eseguito il test d'ipotesi Shapiro-Wilk. Il test ha rigettato l'ipotesi che i campioni relativi ai residui utilizzati siano distribuiti seguendo una distribuzione normale.

7 Conclusioni

Alla luce della fase di valutazione, per raggiungere l'obiettivo di predire la temperatura media del mese successivo, si può concludere che i modelli migliori sono il *Tree Ensemble e Random Forest*, ovvero i due modelli con R^2 maggiore e *MSE* minore, rispetto agli altri modelli. D'altro canto i modelli peggiori per la risoluzione di questo problema sono i modelli di *Linear Regression e Polinomial Regression* che hanno ottenuto i peggiori risultati di R^2 e *MSE*. Un eventuale sviluppo futuro potrebbe consistere in una migliore gestione dei valori mancanti all'interno del dataset, integrando la sorgente dati con altre sorgenti più complete, in modo da avere più dati da poter utilizzare nella fase di apprendimento dei modelli.

References

[1] Kevin P. Murphy, *Machine Leraning A probabilistic Perspective*.

[2] Knime WebSite, `https://www.knime.com/`.

[3] Dataset Climate Change: Earth Surface Temperature Data, `https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature\-data`

[4] John Forbes Kerry.

[5] XGBoost for Regression,
`https://machinelearningmastery.com/xgboost-for-regression/`
Jason Brownlee, 12-02-2021

[6] Coefficient of Determination
`https://en.wikipedia.org/wiki/Coefficient_of_determination`

[7] Mean squared error
`https://en.wikipedia.org/wiki/Mean_squared_error`