Bike Sharing
Report Fondamenti di Machine Learning

Giulia Lui – 308955@studenti.unimore.it

0 - INTRODUZIONE

Il dataset è reperibile tramite questo link:
https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

Il progetto ha come obiettivo la creazione di un modello di Machine Learning per il task di regressione. Lo scopo è prevedere il numero totale di noleggi di biciclette in un sistema di bike-sharing. Il dataset utilizzato contiene informazioni orarie relative al noleggio di biciclette nel periodo compreso tra il 2011 e il 2012.

Il dataset è composto da un totale di 17379 samples e 17 variabili tra cui la variabile target. Ogni riga rappresenta il conteggio orario dei noleggi. La variabile target da prevedere è 'cnt', che corrisponde al conteggio totale dei noleggi, includendo sia gli utenti occasionali ('casual') che quelli registrati ('registered'). Il dataset non presenta valori mancanti.

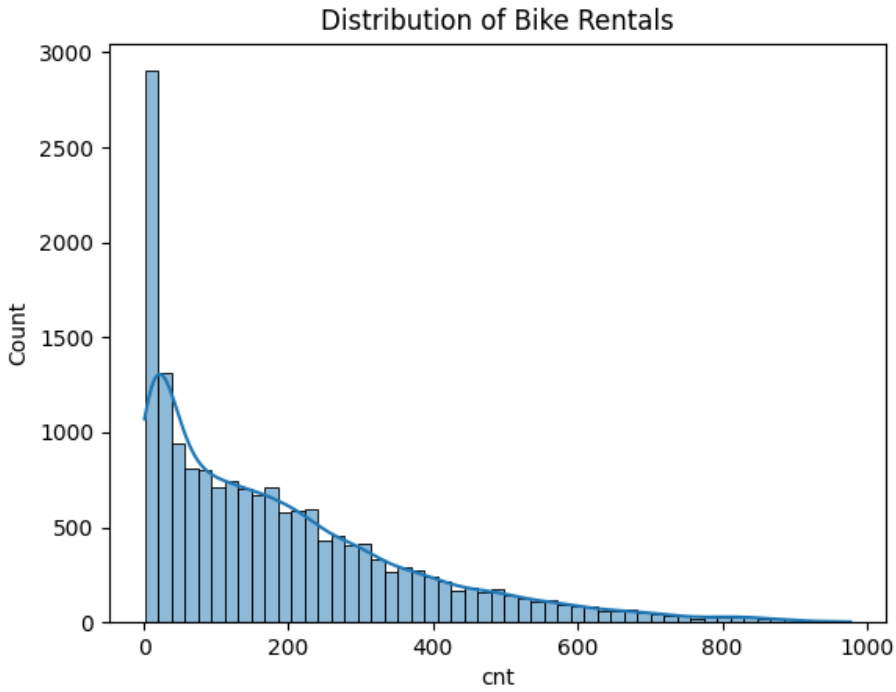
Prima di procedere con l'analisi e la modellazione, sono state rimosse alcune variabili. La variabile 'instant' è stata scartata perché non utile ai fini predittivi. La variabile 'dteday' è stata scartata in quanto le informazioni che si potevano estrarre da essa (giorno della settimana e mese) erano già presenti nel dataset in variabili separate ('weekday' e 'mnth'). Inoltre, le variabili 'casual' e 'registered' sono state eliminate, poiché la variabile target 'cnt' è la somma di queste due variabili e mantenendole, la previsione sarebbe stata banale e poco significativa perché il modello si sarebbe limitato a sommare i loro valori invece di imparare e prevedere il conteggio totale basandosi sulle altre caratteristiche.

1 – EDA: EXPLORATORY DATA ANALYSIS

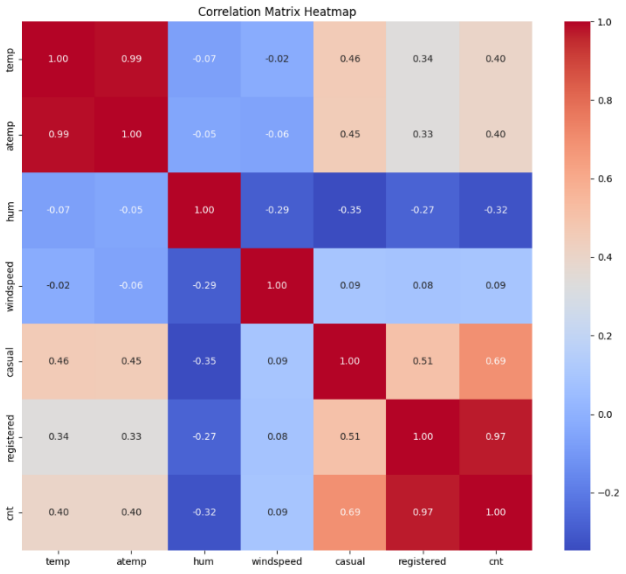
In questa fase iniziale, è stata effettuata un'analisi delle caratteristiche del dataset. Il dataset non presentava valori mancanti o nulli. Per prevedere la variabile target 'cnt', sono state considerate variabili numeriche ('temp', 'atemp', 'hum', 'windspeed'), variabili binarie ('holiday', 'workingday') e variabili categoriche ('season', 'yr', 'mnth', 'hr', 'weekday', 'weathersit'). Poiché i modelli di machine learning richiedono input numerici, le variabili categoriche sono state trasformate utilizzando il **Label Encoding**, che ha assegnato un numero intero ad ogni categoria.

Per comprendere la natura della variabile target, è stato generato un istogramma della sua distribuzione. Il grafico "Distribution of Bike Rentals" mostra una distribuzione non normale,

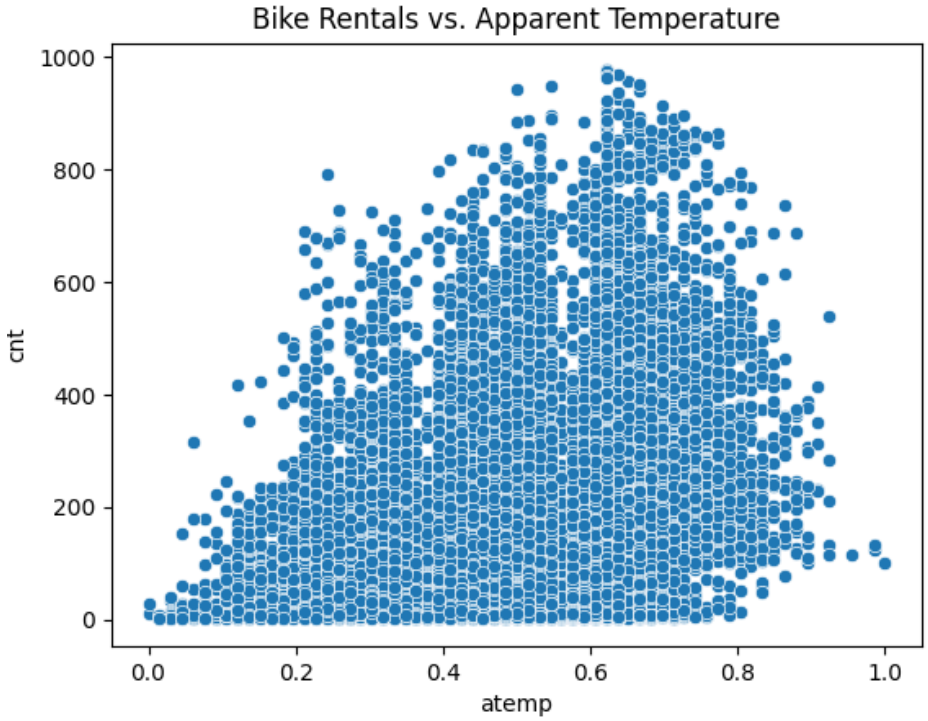
con una notevole asimmetria verso sinistra. La maggior parte dei dati si concentra su un numero di noleggi ridotto, mentre un numero minore di osservazioni corrisponde ad un numero di noleggi elevato.



Per analizzare le relazioni tra le variabili, è stata generata una heatmap della matrice di correlazione. Da essa, ad esempio, si deduce una correlazione molto forte tra 'temp' e 'atemp', essendo feature molto simili tra di loro. A causa di questa forte relazione, la variabile 'temp' è stata rimossa, mantenendo solamente 'atemp', per evitare che le due variabili forniscano informazioni ridondanti al modello.



È stato generato anche un grafico a dispersione “Bike Rentals vs. Apparent Temperature” che evidenzia un’interessante relazione tra la temperatura percepita e il numero di noleggi. All’aumentare della temperatura, anche il numero di noleggi aumenta. Tuttavia, la relazione non è perfettamente lineare, ma mostra una curva che suggerisce un picco di noleggi a temperature intermedie. Questa osservazione conferma l’importanza delle condizioni climatiche come fattore predittivo e giustifica l’utilizzo di modelli di regressione non lineari.



2 – PRE-PROCESSING DEI DATI E SUDDIVISIONE

Per garantire una valutazione imparziale della performance del modello, il dataset è stato diviso in due parti utilizzando una suddivisione standard: un training set (80%) e un testing set (20%).

Per ottimizzare le prestazioni dei modelli è stato applicato un processo di scalamenti dei dati tramite **standardizzazione**. Trasforma i dati in modo che abbiano una media pari a 0 e una deviazione standard pari a 1. Questo scalamento è stato applicato separatamente sul training set per apprendere i parametri e poi sul testing set, per garantire che la fase di test simulasse un’applicazione reale del modello su dati completamente nuovi.

3 – DEFINIZIONE DEI MODELLI E TRAINING

Per trovare il modello di regressione più adatto, sono stati scelti diversi algoritmi con caratteristiche differenti. I modelli selezionati per la fase iniziale sono stati la **Linear**

Regression, un modello base di tipo lineare, e modelli ad albero come il **Decision Tree Regressor** e il **Random Forest Regressor**. La scelta degli ultimi due modelli è stata motivata dall'analisi esplorativa dei dati (EDA) che ha mostrato una relazione non lineare, più adatta a essere catturata da modelli non lineari come gli alberi di decisione.

Per garantire che ogni modello fosse addestrato con parametri ottimali, è stata utilizzata la tecnica di **Grid Search** in combinazione con la **Cross-Validation** (con 5 fold). Questo metodo permette di testare diverse combinazioni di iper-parametri e di selezionare quella che fornisce il punteggio migliore sulla metrica di valutazione scelta ('neg_mean_squared_error': negativo dell'errore quadratico medio).

Per confrontare le prestazioni di ogni modello dopo l'ottimizzazione degli iper-parametri, è stata eseguita una valutazione utilizzando le metriche dell'errore quadratico. Questi risultati preliminari hanno permesso di identificare i modelli più promettenti per la fase successiva.

```
Linear Regression
Best MSE:  20334.26
Best RMSE: 142.60

Decision Tree
Best MSE:  3670.13
Best RMSE: 60.58
The best choice for parameter max_depth:  15

Random Forest
Best MSE:  1966.21
Best RMSE: 44.34
The best choice for parameter n_estimators:  200
The best choice for parameter max_depth:  15
```

Per migliorare ulteriormente le prestazioni, è stato creato il modello ensemble **Stacking Regressor**. Questo modello combina le previsioni dei modelli di base (Linear Regression, Decision Tree Regressor e Random Forest Regressor) per ottenere una previsione finale più robusta e accurata. Si può notare nella figura sotto come l'errore sia ulteriormente diminuito.

```
--- Stacking Regressor ---
The cross-validated RMSE of the Stacking Ensemble meta-model is  44.15
```

4 – TESTING E VALUTAZIONE DELLA PERFORMANCE FINALE

Dopo aver addestrato il modello finale (Stacking Regressor) sui dati di training, è stata eseguita una valutazione sul set di test, composto dal 20% dei dati originali. La valutazione è

stata condotta utilizzando le metriche di regressione: **Errore Quadratico Medio (MSE)**, **Radice dell'Errore Quadratico Medio (RMSE)** e **Coefficiente di Determinazione (R^2).**

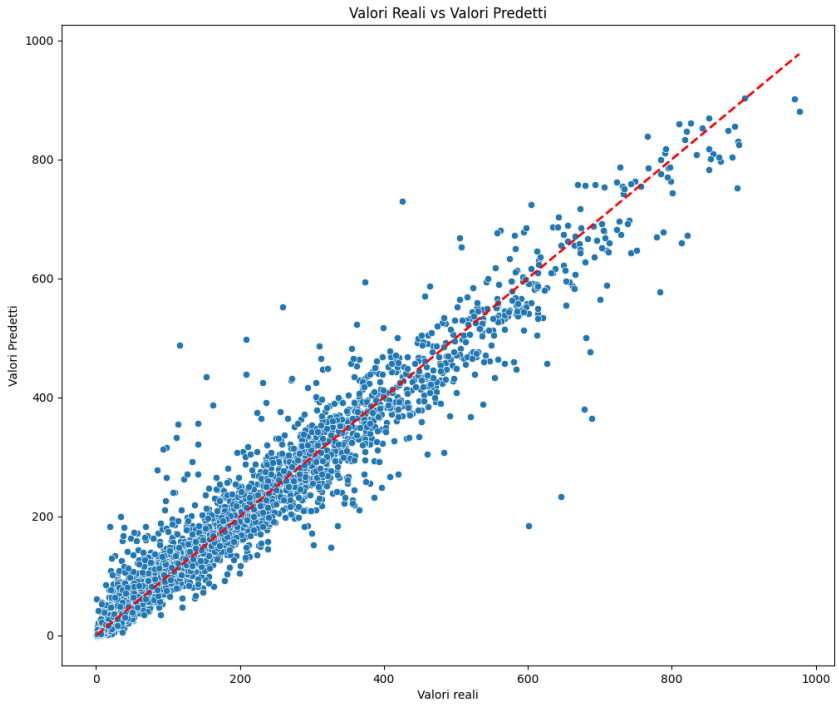
I risultati finali si possono controllare nella figura sotto.

```
---- Final Testing Results ----
Mean Squared Error (MSE):  1757.28
Root Mean Squared Error (RMSE):  41.92
R-squared (R2) Score:  0.94
```

Il valore di R^2 pari a 0,94 indica che il modello è in grado di spiegare il 94% della varianza nel numero di noleggi di biciclette. Questo risultato, insieme ai valori bassi di MSE e RMSE, dimostra una performance elevata e un'ottima accuratezza predittiva.

5 – CONCLUSIONI

Per una valutazione visiva della performance, è stato generato un grafico a dispersione che confronta i valori dei noleggi reali ('y_test') con i valori predetti dal modello ('y_pred'). La linea diagonale rossa rappresenta il caso ideale in cui i valori predetti corrispondono esattamente ai valori reali. Si può notare che i punti si raggruppano intorno a questa linea rossa confermando l'accuratezza del modello.



Dal punto di vista pratico, i risultati ottenuti possono essere utili a chi gestisce servizi di bike-sharing. La possibilità di prevedere la domanda in anticipo permette di migliorare la distribuzione delle biciclette nelle stazioni e la soddisfazione degli utenti, supportando strategie di mobilità sostenibile nelle aree urbane.

Tuttavia, il modello ha delle limitazioni. Le previsioni si basano esclusivamente sulle variabili contenute nel dataset e non considerano fattori esterni come eventi cittadini o fenomeni metereologici estremi, che possono influenzare significativamente l'utilizzo delle biciclette.

Per sviluppi futuri, si potrebbe valutare l'integrazione di nuove feature e l'impiego di modelli più complessi come le reti neurali per apprendere meglio rappresentazioni più complesse e non lineari dei dati.

In sintesi, il progetto ha mostrato come l'applicazione di tecniche di Machine Learning e una fase di pre-processing e selezione dei modelli consenta di ottenere risultati accurati e applicabili in contesti reali.