A.D. 1308
unipg
DIPARTIMENTO
DI INGEGNERIA

Progetto di

Data Intensive Application and Big Data

Corso di Laurea in Ingegneria Informatica e Robotica

Curriculum Data Science – A.A. 2023-2024

Dipartimento di Ingegneria

docente

Prof. Fabrizio Montecchiani

House Price Prediction

363433 **Gian Marco Ferri** gianmarco.ferri@studenti.unipg.it

Link GitHub: `https://github.com/Gian99Marco/House-Price-Prediction`

1 Introduzione

Questo progetto mira a ottenere predizioni sui prezzi delle case utilizzando un modello di regressione lineare implementato attraverso la libreria MLlib di Apache Spark. L'applicazione sfrutta questo modello, addestrato su un dataset di caratteristiche immobiliari, per stimare il prezzo di una casa in base a specifiche caratteristiche fornite dall'utente tramite linea di comando. L'obiettivo principale è fornire uno strumento semplice e rapido per stimare il valore di mercato di una proprietà.

2 Data Flow e tecnologie utilizzate

2.1 Dataset e formato di input

Il dataset utilizzato per l'addestramento del modello include le seguenti informazioni:

- Qualità complessiva
- Area abitabile (in piedi quadrati)
- Numero di posti auto nel garage
- Area del garage (in piedi quadrati)
- Area totale del seminterrato (in piedi quadrati)
- Numero di bagni completi
- Anno di costruzione
- Numero di camere da letto

Il dataset è reperibile sul sito web Kaggle a questo link: House Prices

2.2 Tecnologie utilizzate

Nel progetto sono state utilizzate le seguenti tecnologie:

- **Apache Spark**: Per l'elaborazione distribuita dei dati e l'addestramento del modello di machine learning.
- **MLlib**: libreria di Apache Spark per il machine learning.
- **Java**: Il linguaggio di programmazione utilizzato per implementare il progetto.
- **Maven**: Per la compilazione del progetto e la gestione delle dipendenze.

2.3 Architettura e dataflow

1. **Caricamento dei dati**: Il dataset viene caricato da un file CSV in un DataFrame di Spark.
2. **Preprocessing dei dati**: I dati vengono puliti e preprocessati, gestendo i valori mancanti e trasformando le feature se necessario.
3. **Feature Engineering**: Le feature rilevanti vengono selezionate e assemblate in un vettore di feature.
4. **Addestramento del modello**: Viene creato un modello di regressione lineare e aggiunto ad una pipeline di Spark, che viene poi addestrata utilizzando i dati preprocessati
5. **Predizione**: Il modello viene utilizzato per fare previsioni sui dati inseriti in input dall'utente.

3 Use case

Questo progetto può essere utilizzato da agenzie immobiliari, analisti finanziari e acquirenti o venditori individuali per stimare il prezzo di una casa basandosi sulle sue caratteristiche, aiutando a prendere decisioni riguardo all'acquisto o alla vendita di immobili.

3.1 Input

All'avvio del programma, all'utente vengono richiesti tramite linea di comando diversi parametri relativi alla casa per la quale si vuole stimare il prezzo. I parametri richiesti sono:

- Qualità complessiva (*Overall Quality*) su una scala da 1 a 10.
- Superficie abitabile sopra il livello del suolo (*Above Ground Living Area*) in piedi quadrati (*sq ft*).
- Numero di posti auto nel garage (*Number of Garage Cars*).
- Superficie del garage (*Garage Area*) in piedi quadrati (*sq ft*).
- Superficie totale del seminterrato (*Total Basement Area*) in piedi quadrati (*sq ft*).
- Numero di bagni completi (*Number of Full Bathrooms*).

- Anno di costruzione (*Year Built*).
- Numero di camere da letto sopra il livello del suolo (*Number of Bedrooms Above Ground*).

Un esempio di input è mostrato di seguito:

```
Enter Overall Quality (1-10): 8
Enter Above Ground Living Area (sq ft): 220
Enter Number of Garage Cars: 2
Enter Garage Area (sq ft): 22
Enter Total Basement Area (sq ft): 30
Enter Number of Full Bathrooms: 3
Enter Year Built: 2001
Enter Number of Bedrooms Above Ground: 3
```

3.2 Output

Dopo aver inserito i dati, il programma utilizza il modello di regressione lineare per prevedere il prezzo della casa. Il risultato viene quindi visualizzato all'interno di un rettangolo di asterischi, con due cifre decimali, per una migliore leggibilità.
Un esempio di output è mostrato di seguito:

```
************************************
* Predicted House Price: $111775,94 *
************************************
```

Questo formato garantisce che il prezzo predetto sia chiaramente visibile e facilmente distinguibile.

4 Limitazioni e Possibili Estensioni

4.1 Limitazioni

- La validità del modello dipende dai dati su cui è addestrato: se i dati non fossero rappresentativi dello scenario reale, le previsioni potrebbero non essere accurate.
- Il modello non tiene conto dei fattori economici, delle caratteristiche della località o di altre variabili esterne che possono influenzare i prezzi delle case.

4.2 Possibili estensioni

- Integrare ulteriori caratteristiche come i tassi di criminalità del quartiere, le valutazioni delle scuole, la vicinanza ai servizi, ecc.

- Sperimentare con altri algoritmi di machine learning come Decision Trees, Random Forests o Gradient Boosting.
- Implementare un'interfaccia web per ottenere una maggiore accessibilità e interazione con l'utente.