



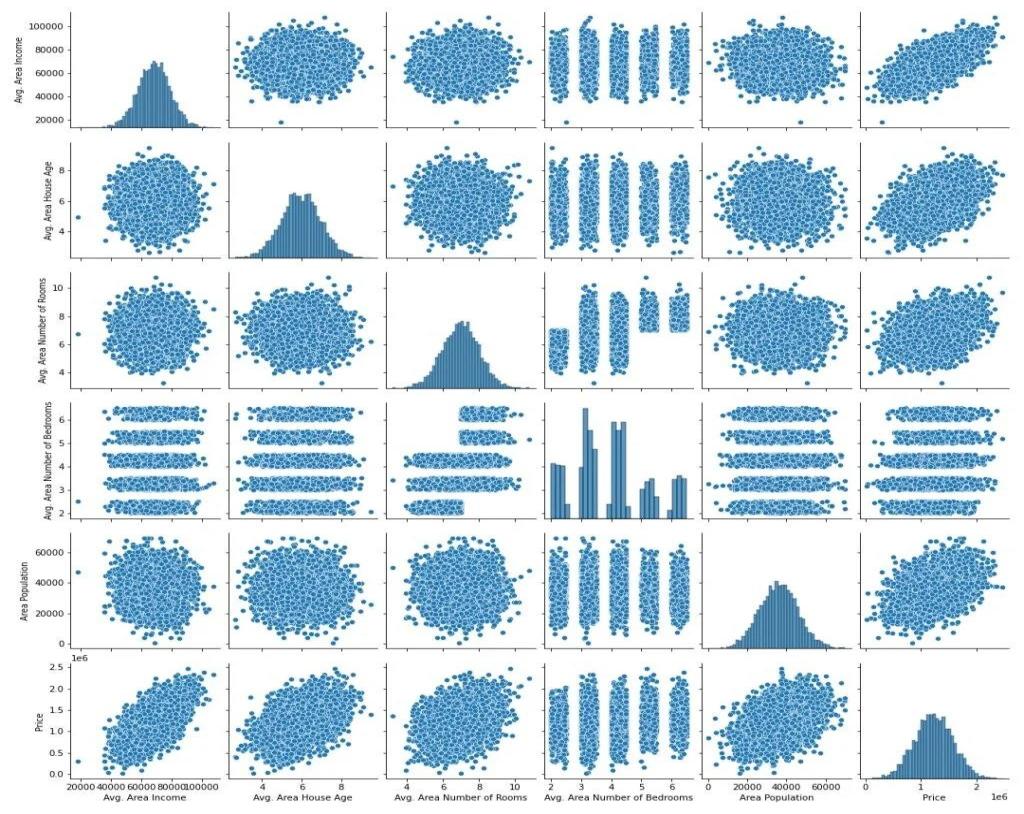


**What is machine learning?**

•Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/topics/artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Problem definition

* Prices of real estate properties are sophisticatedly linked with our economy.
* Despite this, we do not have accurate measure of house prices based on the vast amount of data available.
* Proper and justify prices of properties can bring in a lot transparency and trust back to real estate industry, which is very important for most consumers especially in India



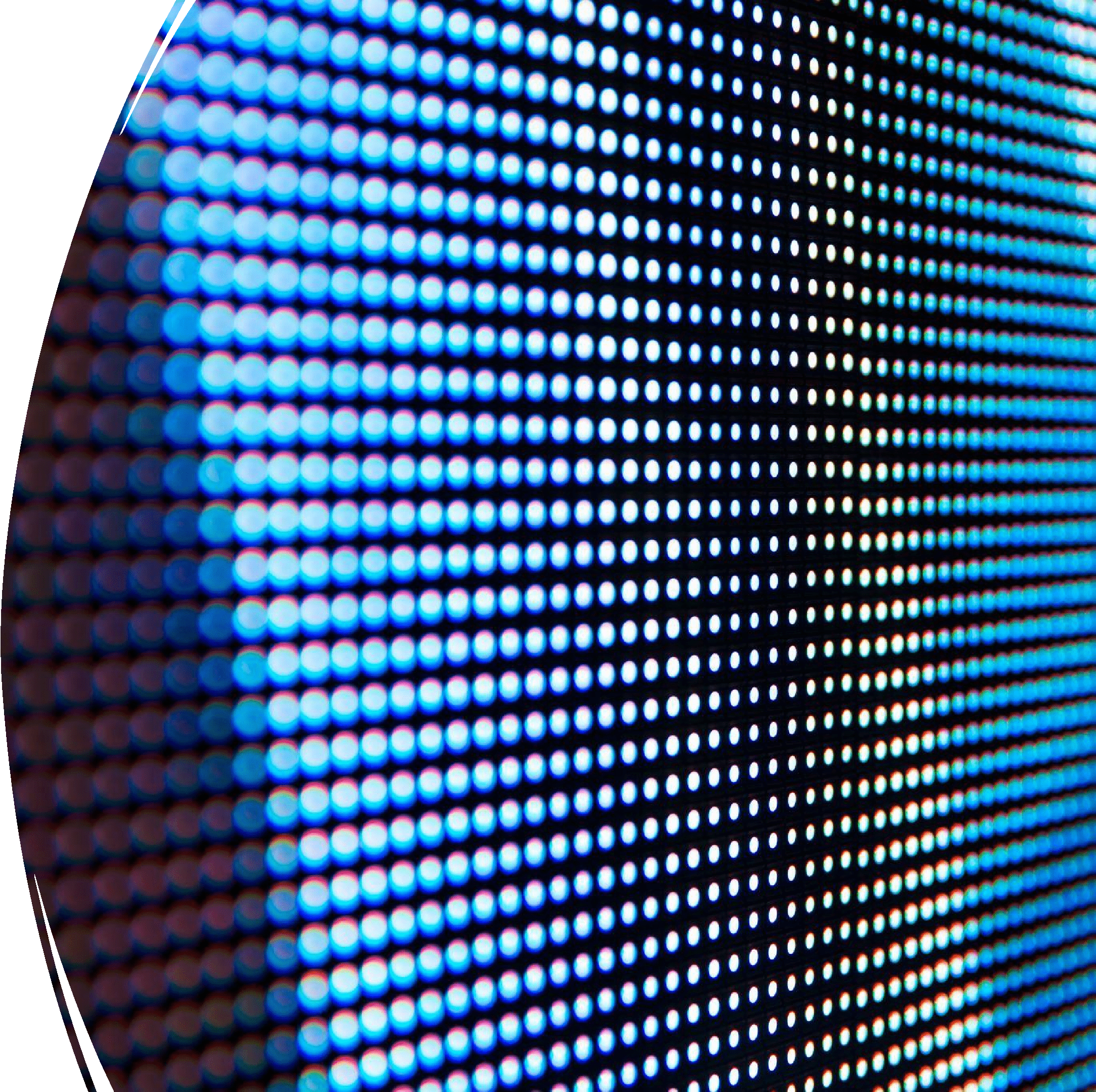
**Step 7 – Splitting our data for training and test**

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=101)

**Step 10 – Plotting the residuals of our House Price Prediction model.**

sns.histplot((y\_testpred),bins=50,kde=True)

**Step 11 – Observe the coefficients.**

cdf=pd.DataFrame(lr.coef\_, cols, ['coefficients']).sort\_values('coeffici ents',ascending= False ) cdf

**Data preprocessing for House Price Prediction:**

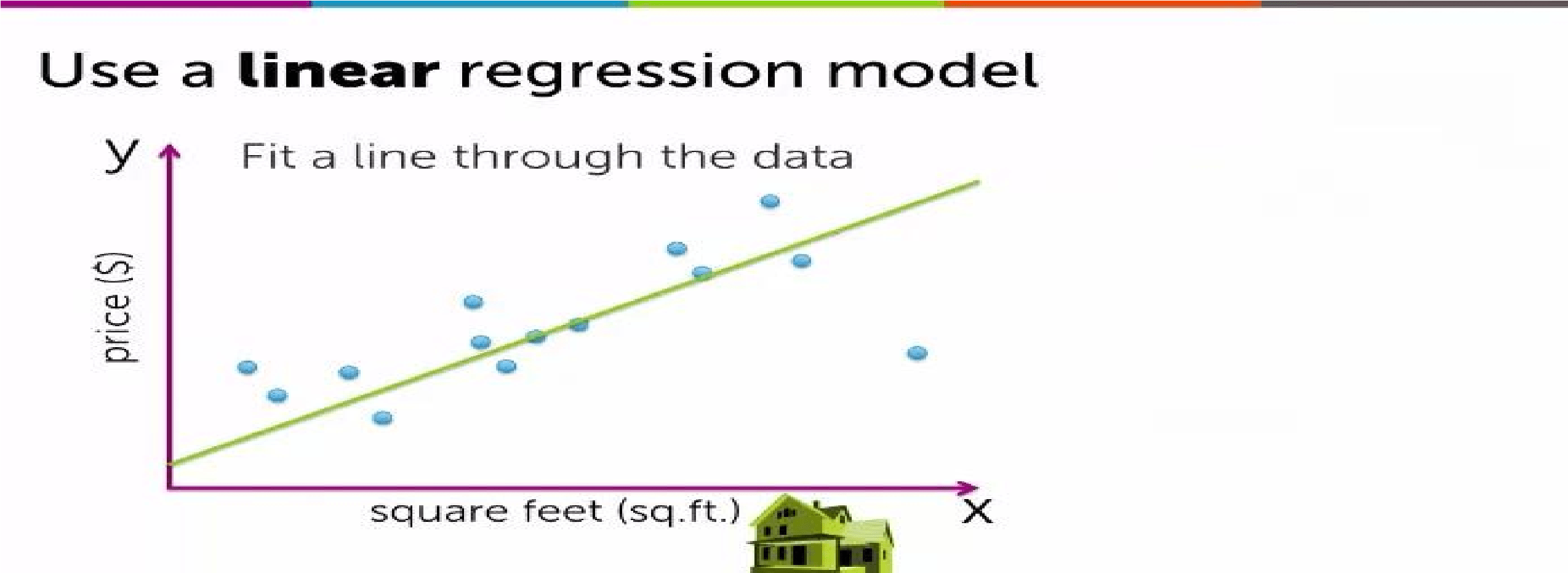
**Data Preprocessing:**

Data preprocessing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, greater is the reliance on the produced results. **Incomplete, noisy, and inconsistent data** are the properties of large real-world datasets. Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise and resolving inconsistencies.

* **Incomplete data** can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions.
* There are many possible reasons for **noisy data** (having incorrect attribute values). The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Errors in data transmission can also occur. Incorrect data may also result from inconsistencies in naming conventions or data codes used, or inconsistent formats for input fields, such as date.

There are a number of data preprocessing techniques available such as,

* **Data Cleaning**
* **Data Integration**
* **Data Transformation**
* **Data Reduction**



Training a Linear Regression

Model

To train the linear regression model, a training set is used. The training set comprises pairs of **input features***(eg: square feet)* and corresponding **output targets** *(eg: house price)*.

Let:

* *“m” be the number of training examples.*
* *“x” be the input feature representing a specific property of the house, such as its size.*
* *“y” be the output target, representing the actual price of the house.*
* *“i” be the index of a specific training example.* A specific training example is denoted as (x^i, y^i).

Therefore, the equation can be written as:

Training set: {(x¹, y¹), (x², y²),

…, (x^m, y^m)}

*where each pair* ***(x^i, y^i)*** *represents a specific house in the dataset, and “m” denotes the number of training examples.*

**The Function f(x) and Model Representation:**

When training the model, a learning algorithm produces a function (f) that takes **an input feature (x)** and **outputs an estimate or prediction (**ŷ**)**. In linear regression, the function f(x) is defined as:

*where:*

•*“w” and “b” are numeric values that determine the line’s* ***slope*** *and* ***intercept****, respectively*

A crucial aspect of linear regression is the cost function. The cost function measures the model’s performance by quantifying the difference between predicted values and actual targets. This concept is widely applicable in machine learning and plays a vital role in training advanced AI models.

In the next article, we’ll delve deeper into the cost function and explore how it contributes to training linear regression models.

In conclusion, linear regression is a valuable tool for predicting house prices based on specific features. Its simplicity and effectiveness make it widely used in the field of machine learning. By fitting a straight line to the data, the linear regression model can estimate the price of a house based on its size. This example demonstrates the concept of supervised learning, where the model is trained using labeled data to predict numeric outputs.



. absolute error (MAE), Mean squared error (MSE), Median absolute error (MedAE) and Coefficient of determination (R2 ). They are all defined below.

**Mean absolute error (MAE)**

Mean absolute error measures the prediction error by taking the mean of all absolute values of all errors, that is: MAE = Pn i=0 |yi − yˆi | n

Where n is the number of samples, y are the target values and yˆ are the predicted values. A MAE closer to 0 means that the model predicts with lower error and that the prediction is better the closer the MAE is to 0.

**Mean squared error (MSE)**

Mean squared error is similar to MAE, but the impact of a term is quadratically proportional to its size. It measures the prediction error by taking the mean of all squared absolute values of all errors, that is:

MSE= Pn i=0(yi − yˆi)

2 n METHODS As for MAE, values close to 0 are better.

**Median absolute error (MedAE)**

The median absolute error (MedAE) is the median of all absolute differences between the predicted value and the target value. In difference to MAE and MSE, the median absolute error is more robust to outliers by virtue of using the median instead of the mean.

MedAE = median(|y1 − yˆ1|...|yn − yˆn|)

A low MedAE means little error and a good prediction.

**Coefficient of determination (R2 )**

The coefficient of determination, R2 , is the ratio between the explained variance and the total variance. Another, perhaps more intuitive way of understanding it is that it is the fraction of the variance that is predictable by (or explained by) by the model. It is frequently used to evaluate how well a regression model fits the data. The coefficient value ranges from 0 to 1 where 1 is better, meaning perfect determination and 0 meaning no determination. In this study, it is calculated using the scikit-learn r2\_score function as follows .

R 2 (y, yˆ) = 1 − Pnsamples−1 i=0 (yi − yˆi) 2 Pnsamples−1 i=0 (yi − y¯) 2

Where yˆi is the predicted value of the sample, y is the corresponding actual value over n samples and y¯ is the arithmetic mean as such: y¯ = 1 n samples n samples X−1 i=0 yi

**Conclusion**

In conclusion, using machine learning in Python is a powerful tool for predicting house prices. By gathering and cleaning data, visualizing patterns, and training and evaluating our models, we can make informed decisions in the dynamic world of real estate.

By leveraging advanced algorithms and data analysis, we can make accurate predictions and inform decision-making processes. This approach empowers buyers, sellers, and investors to make informed choices in a dynamic and competitive market, ultimately maximizing their opportunities and outcomes.