Applied AI & ML Industry Projects Lab (AAM-IPL)
Transforming Learning into Industry Solutions

Week-3 – Linear Regression

Linear Regression and Boston Housing Price Prediction Project Overview

V Semester - ML and AIUP, Aug-Oct 2024

Session Date and Time: 28th Sept 2024, 10:30 AM IST – 12:00 Noon IST

Venkateswar Reddy Melachervu
Visiting Faculty and CTO, Brillium Technologies
Department of Computer Science Engineering – AI and ML (CSM)
Email: venkat.reddy.gf@gprec.ac.in

G.Pulla Reddy Engineering College (Autonomous)
G.Pulla Reddy Nagar, Nandyal Road, Kurnool, AP 518007, India
Website: https://www.gprec.ac.in

Disclaimer and Confidentiality Notice

The content of this guest lecture, including all discussions, materials, and demonstrations, code is intended for educational purposes only and reflects the views and opinions of the speaker. While every effort has been made to ensure the accuracy and relevance of the information presented, it should not be considered as legal, financial, or professional advice.

Brillium Technologies retains unrestricted ownership of all information shared during this session. Participants must not record, reproduce, distribute, or disclose any part of the lecture or materials without prior written permission from Brillium Technologies. Unauthorized use or distribution of the content may result in legal action.

Additionally, all trademarks, service marks, and intellectual properties referenced or used in this presentation are the property of their respective owners. No ownership or rights over such third-party content are claimed or implied by the author or Brillium Technologies or by GPREC.

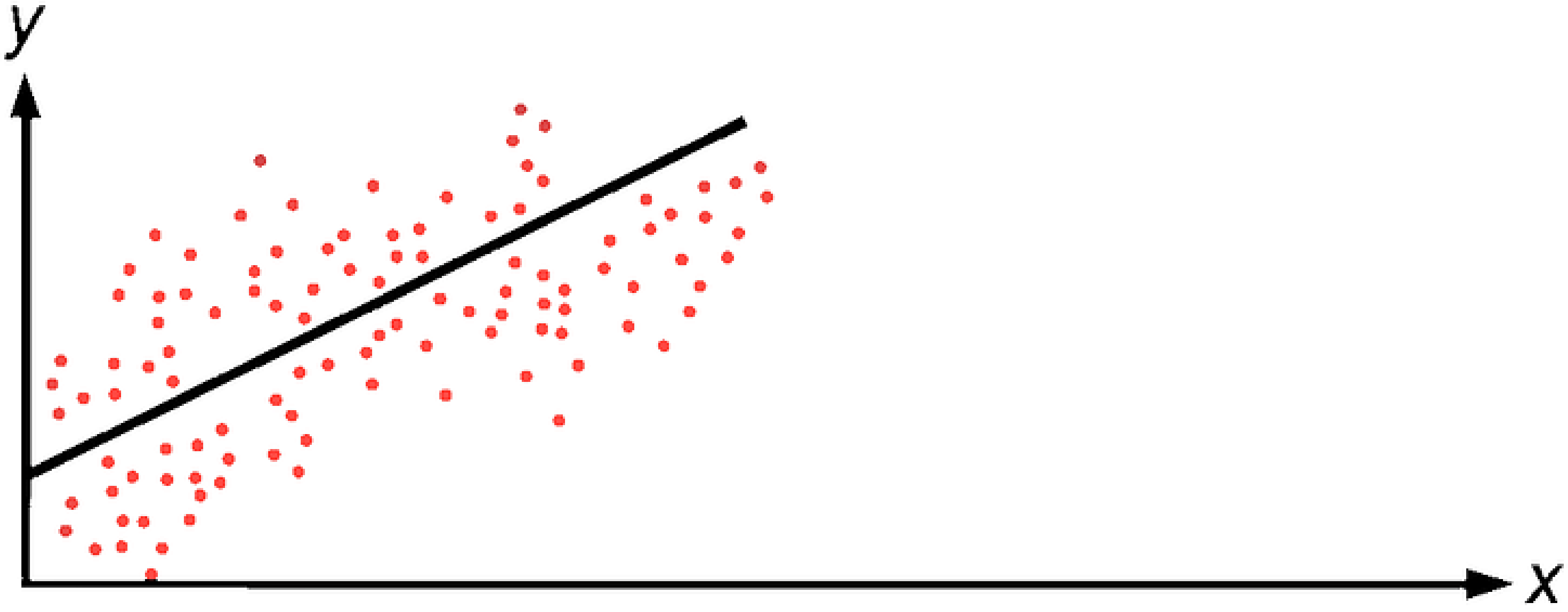
By attending this lecture, you agree to respect the confidentiality of the information shared and refrain from using it in any unauthorized manner. Failure to comply with these terms may result in legal action.

Thank you for your understanding and cooperation.

Lecture Outline

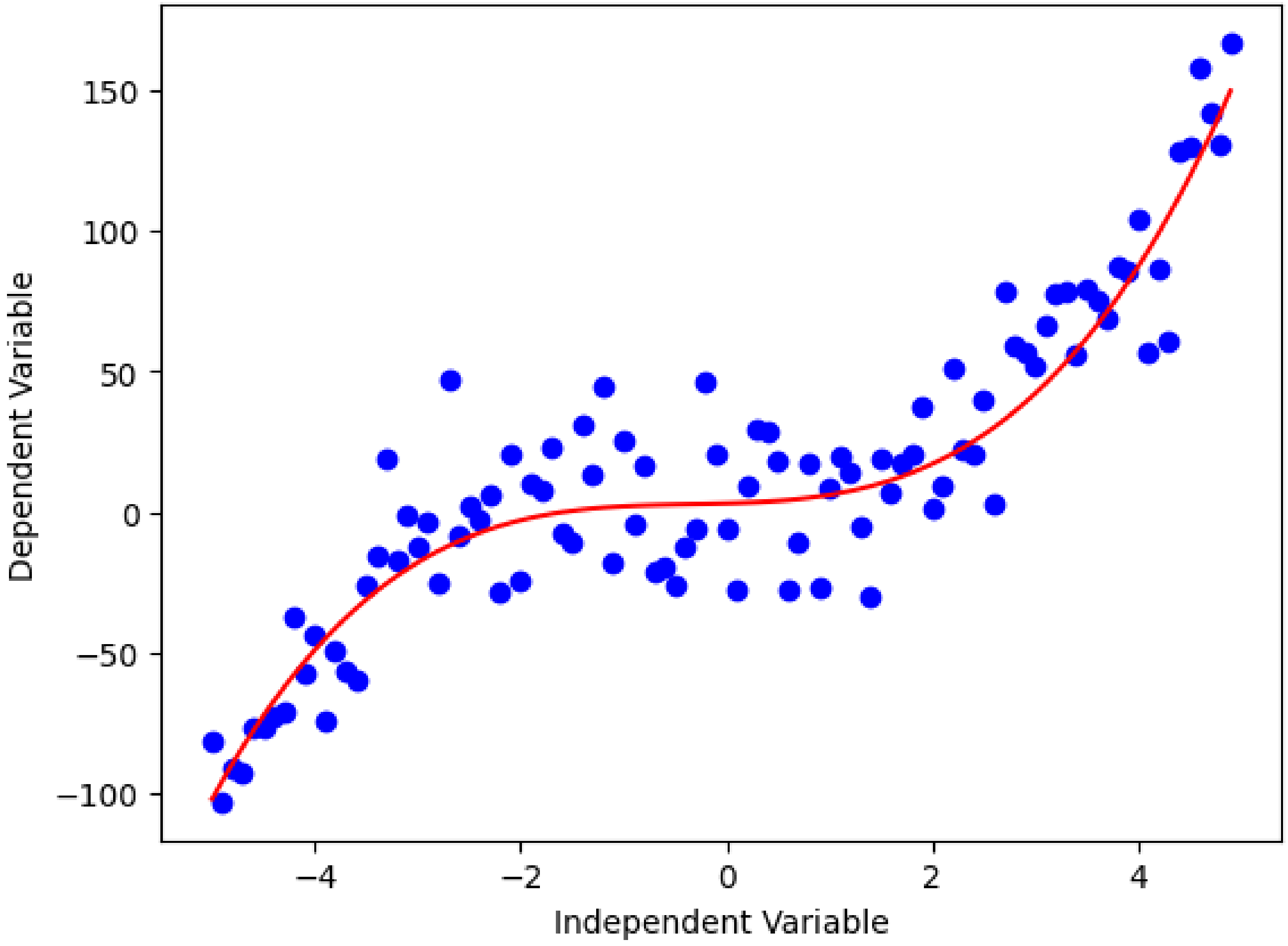
- Data Sets and Distribution
- Data Points and Line of Regression
- What is Linear Regression?
- Dependent and Independent Variables
- Linear Regression Visualisation
- Mathematical Model and Training
- Computing Regression Coefficients
- Boston Housing Price Prediction Project - Overview and Implementation Steps
- Demo
- Q&A

Data Sets and Distribution

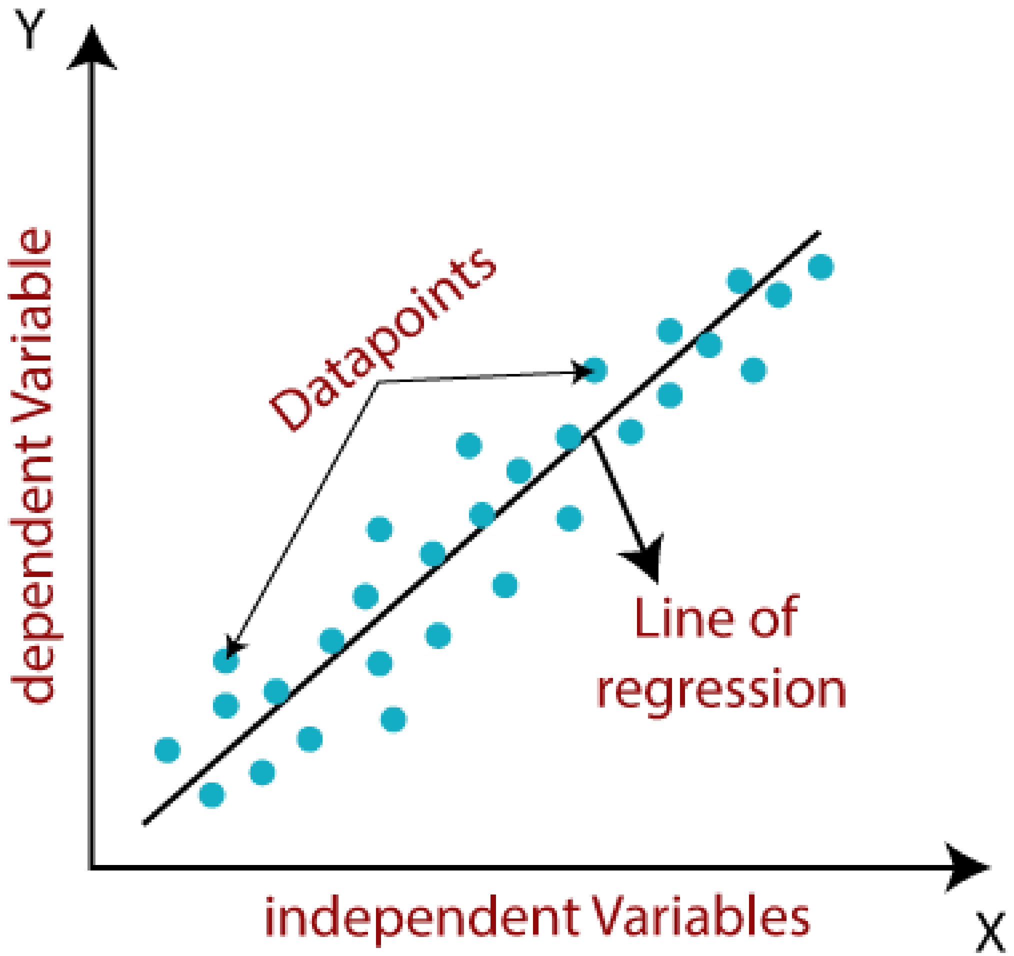


- Y is plotted based on X
- y is dependent or related on $x \Rightarrow y = f(x)$

Data Sets and Distribution



Data Points and Line of Regression



What is Linear Regression?

- A statistical method
- Models relationships between dependent and independent variables for estimating the dependent variable
- Dependent variable is called **target variable** and independent variables are called regressors
- Examples
    - Predicting stock prices, weather forecasting, housing prices, etc.

Linear Regression

- Types of regressions
  - Simple – univariate regression
    - One dependent variable and one independent variable
  - Multiple – multi-variate regression
    - One dependent variable and multiple independent variable

Dependent and Independent Variables

Regressor $\overline{x}$ or Independent Variable of 13-dimensions

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
...														

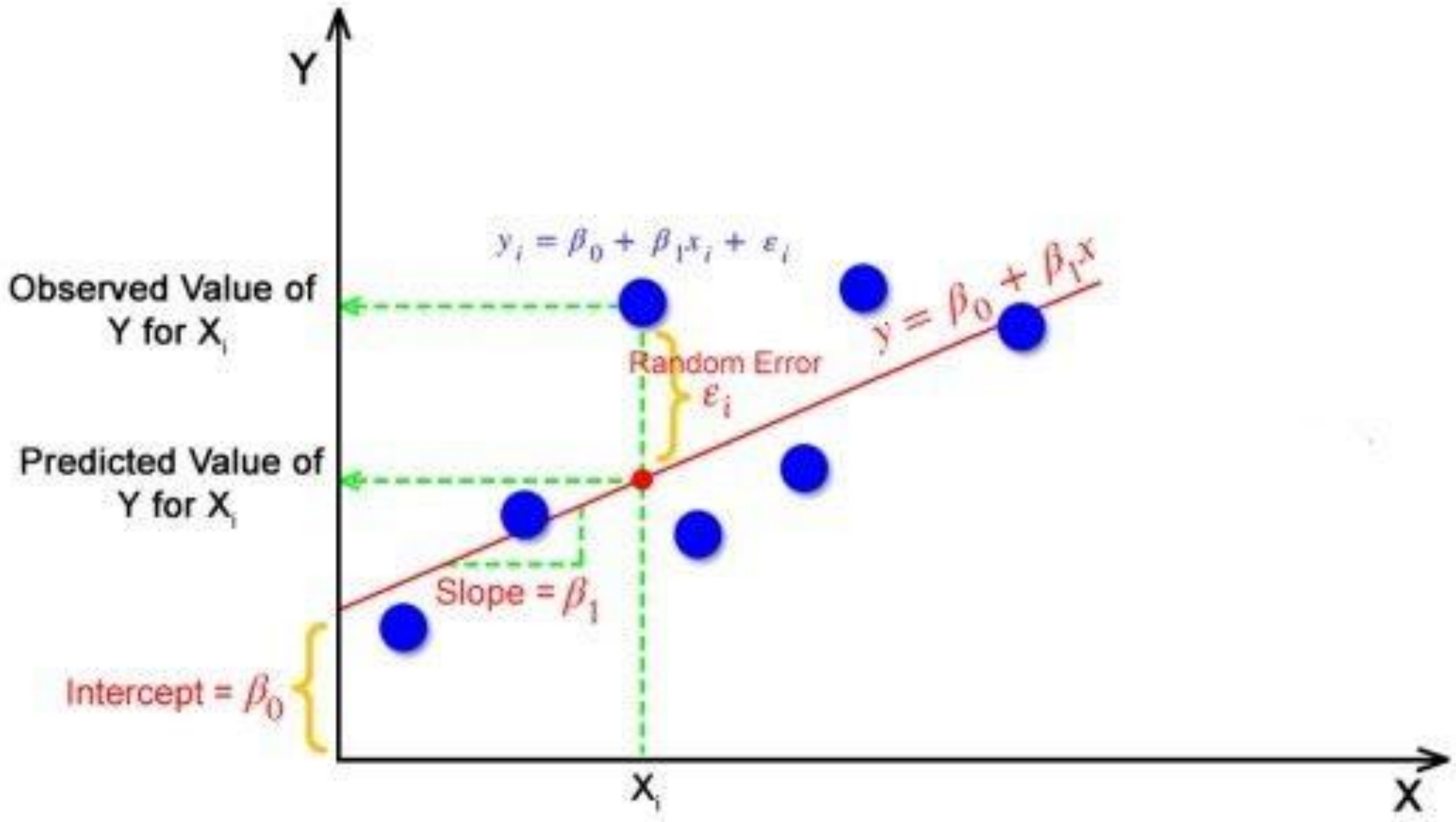
[506 rows x 14 columns]

Response or Target or Dependent Variable

CRIM Per capita crime rate by town
ZN Proportion of residential land zoned for lots over 25,000 sq. ft.
INDUS Proportion of non-retail business acres per town
CHAS Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX Nitric oxide concentration (parts per 10 million)
RM Average number of rooms per dwelling
AGE Proportion of owner-occupied units built prior to 1940
DIS Weighted distances to five Boston employment centres
RAD Index of accessibility to radial highways
TAX Full-value property tax rate per $10,000
PTRATIO Pupil-teacher ratio by town
B 1000(Bk - 0.63)^2 where Bk is the proportion of Black residents by town
LSTAT Percentage of lower-status population
MEDV Median value of owner-occupied homes in $1000s (target variable)

$$\overline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{13} \end{bmatrix}$$

Linear Regression Visualisation



Mathematical Model

- In Boston housing data set, housing price is captured as a function of 13 other independent features/variables
  - Number of rooms, per capita crime in the town, distance to Boston employment center, Charles river bank side, etc.
- Price of any given house $k, y(k)$ in Boston with above independent feature values $\bar{x}_i$ is:

$$y(k) = h_0 \times 1 + h_1 x_1(k) + h_2 x_2(k) + h_3 x_3(k) + \cdots + h_{13} x_{13}(k) + \epsilon(k)$$

Bias

Regression Coefficients - Weights

Model Error

Regressors

Mathematical Model

$$y(k) = h_0 \times 1 + h_1 x_1(k) + h_2 x_2(k) + h_3 x_3(k) + \cdots + h_n x_n(k) + \epsilon(k)$$

$$y(k) = \underbrace{\begin{bmatrix} 1 & x_1(k) & x_2(k) & x_3(k) & \dots & x_n(k) \end{bmatrix}}_{\bar{x}^T} \underbrace{\begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}}_{\bar{h}} + \underbrace{\begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{n+1} \end{bmatrix}}_{\bar{\epsilon}}$$

$$y(k) = \bar{x}^T(k) \times \bar{h} + \epsilon(k)$$

Linear Regression

Training

- Finding the values of regression coefficients using training data set
- Training data set can be mathematically represented as pairs of target and regressor vectors
  - $\left(y(k), \bar{x}(k)\right), k = 1\ to\ M$

$$\begin{bmatrix} y(1) \\ y(2) \\ y(3) \\ \vdots \\ y(M) \end{bmatrix} = \begin{bmatrix} \bar{x}^T(1) \\ \bar{x}^T(2) \\ \bar{x}^T(3) \\ \vdots \\ \bar{x}^T(M) \end{bmatrix} \bar{h} + \begin{bmatrix} \epsilon(1) \\ \epsilon(2) \\ \epsilon(3) \\ \vdots \\ \epsilon(M) \end{bmatrix}$$

$\bar{y}\ - M \times 1$ | $X\ - M \times (n+1)$ | $\bar{\epsilon}\ - M \times 1$

$$\bar{y} = X\bar{h} + \bar{\epsilon}$$

Computing Regression Coefficients

$$\bar{y} = X\bar{h} + \bar{\epsilon}$$

$$\bar{\epsilon} = \bar{y} - X\bar{h}$$

- Start with random values for $\bar{h}$
- Goal is to reduce $\bar{\epsilon}$ to zero/minimum across the training dataset

$$\Rightarrow min\ \bar{\epsilon} = min\|\bar{\epsilon}\|^2 = min\left\|\bar{y} - X\bar{h}\right\|^2$$

- Least Squares Problem/Solution
- Solved using matrix algebra leading to $\bar{h}$
- $\bar{h} = (X^T X)^{-1} X^T \bar{y}$
- $(X^T X)^{-1} X^T$ - Pseudo Inverse of X
  - $\Rightarrow (X^T X)^{-1} X^T X = I$

Why Do We Square Error For Minimisation?

- Magnify error to penalise wrong predictions
- Avoiding cancellation of negative and positive errors
- Squaring makes it differentiable – critical for gradient descent algo
- Least square solution has a closed form formula using matrix algebra

Boston Housing Data Set and House Price Prediction

- **Objective - To build a model to predict the MEDV price of a house**
- The Boston Housing dataset is a classic dataset used for regression tasks, particularly in the domain of housing price prediction
- Contains information collected by the U.S. Census Service concerning housing in the Boston suburbs
- The dataset has been widely used to illustrate the workings of machine learning algorithms, particularly linear regression
- Dataset Overview
    - **Number of Samples**: 506
    - **Number of Features Per Sample**: 13
    - **Target Variable**: MEDV (Median value of owner-occupied homes in $1000s)
- Features
    - CRIM: Per capita crime rate by town
    - ZN: Proportion of residential land zoned for lots over 25,000 sq. ft.
    - INDUS: Proportion of non-retail business acres per town
    - CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
    - NOX: Nitric oxide concentration (parts per 10 million)
    - RM: Average number of rooms per dwelling
    - AGE: Proportion of owner-occupied units built prior to 1940
    - DIS: Weighted distances to five Boston employment centres
    - RAD: Index of accessibility to radial highways
    - TAX: Full-value property tax rate per $10,000
    - PTRATIO: Pupil-teacher ratio by town
    - B: $1000(Bk - 0.63)^2$ where Bk is the proportion of Black residents by town
    - LSTAT: Percentage of lower-status population

Implementation Steps

- Perform Exploratory Data Analysis (EDA)
  - Print any missing values in the provided dataset
  - Print total data samples/points in the data set
  - Print first 5 rows of the data in the set
  - Plot histograms of all features with continuous values
  - Plot correlation heatmap of features
- Build a Linear Regression Model
  - Standardize the dataset and train the model (test size – 20%)
  - Predict the target variable (MEDV) using the independent features
  - Plot Actual Vs Predicted Home Prices
- Evaluate the Model - Use evaluation metrics to assess model's performance
  - R^2 Score
  - Mean Squared Error (MSE)
  - Root MSE
  - Mean Absolute Error (MAE)
  - Mean Absolute Percentage Error (MAPE)
- Print the Regression Coefficients of the Model
- Generate the PDF of Code and Output

Model Evaluation Metrics

- R^2 Score (Coefficient of Determination)
  - Measures the proportion of variance in the dependent variable that is predictable from the independent variables. An R2R 2 value of 1 indicates a perfect fit, while 0 indicates no predictive power.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- Mean Squared Error (MSE)
  - Measures the average of the squared differences between actual and predicted values. It's sensitive to outliers due to squaring.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE)
  - It's the square root of MSE, providing an error value in the same units as the dependent variable.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Model Evaluation Metrics

- Mean Absolute Error (MAE)
  - Measures the average of the absolute differences between actual and predicted values. It's less sensitive to outliers compared to MSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Absolute Percentage Error (MAPE)
  - Measures the average of the absolute percentage differences between actual and predicted values.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

Project Implementation Demo


Q & A

Interested in building a Gen AI application?

Reach out to venkat@brillium.in

THANK YOU!

AAML-IPL Brought You in Partnership with:

Brillium Technologies
Sector 7, HSR Layout, Bengaluru 560102, Karnataka, India
Website: www.brillium.in | Email: connect@brillium.in