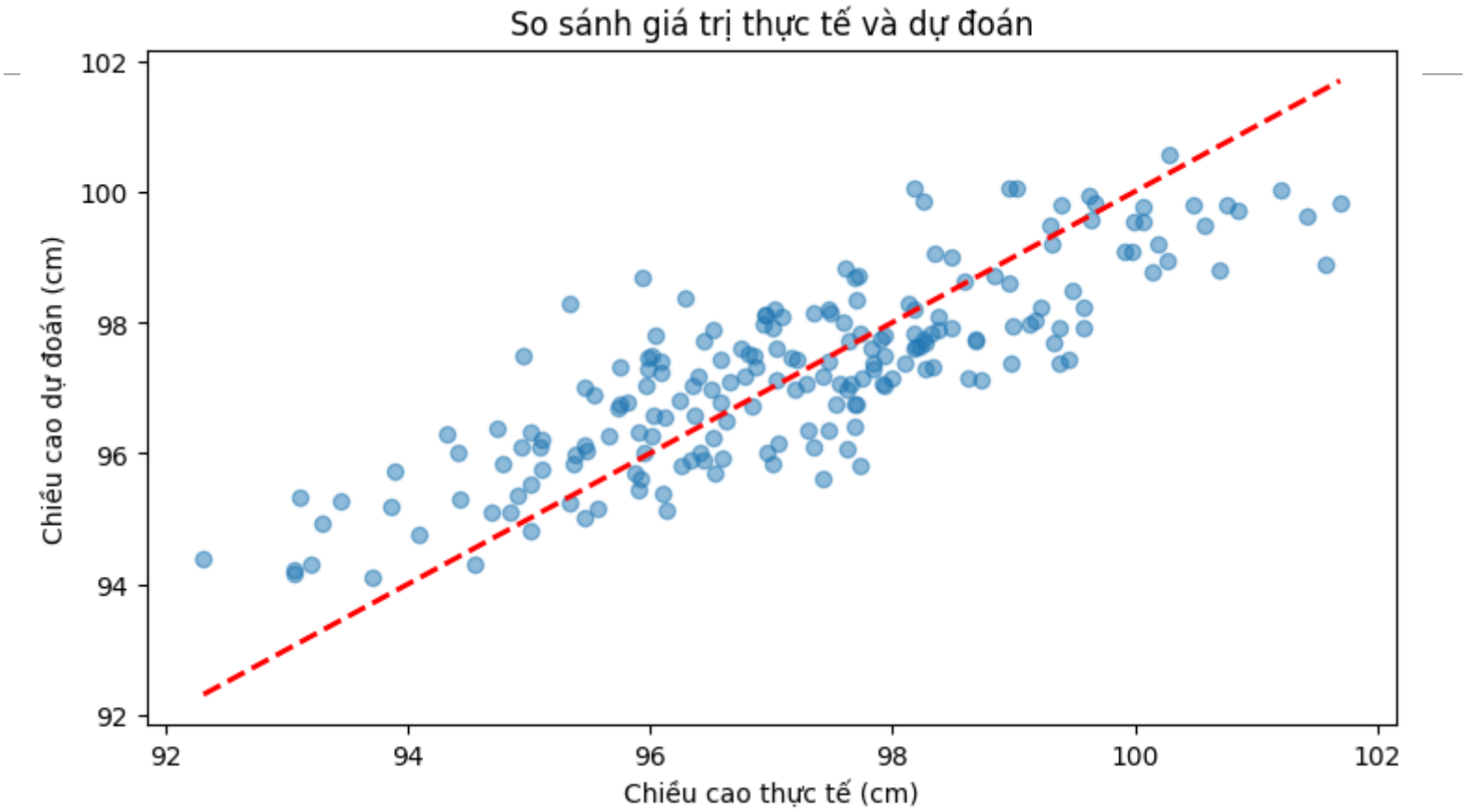
HỌC SÂU

BÀI 2. HỒI QUY TUYẾN TÍNH VÀ HỒI QUY LOGISTIC

Hồi quy tuyến tính

Linear Regression là một thuật toán **học có giám sát (supervised learning)** trong Machine Learning, nó là một phương pháp thống kê dùng để ước lượng mối quan hệ giữa các biến độc lập (input features) và biến phụ thuộc (output target).

Linear Regression giả định rằng sự tương quan giữa các biến là tuyến tính, từ đó tìm ra hàm tuyến tính tốt nhất để biểu diễn mối quan hệ này. Thuật toán này dự báo giá trị của biến output từ các giá trị của các biến đầu vào.


So sánh giá trị thực tế và dự đoán
Chiều cao dự đoán (cm)
Chiều cao thực tế (cm)
92
94
96
98
100
102

Đặc điểm của mô hình

Mục tiêu: Hồi quy tuyến tính hướng đến mô hình hóa và phân tích mối quan hệ giữa một biến phụ thuộc (hay biến mục tiêu) và một hoặc nhiều biến độc lập (hay biến giải thích).

Mục đích của mô hình

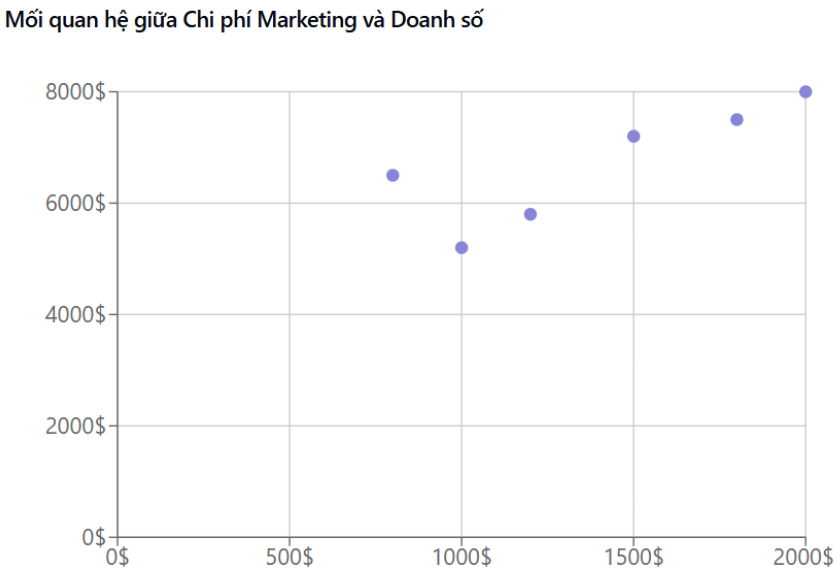
Dự đoán và dự báo: Sử dụng biến độc lập để dự đoán giá trị của biến phụ thuộc.

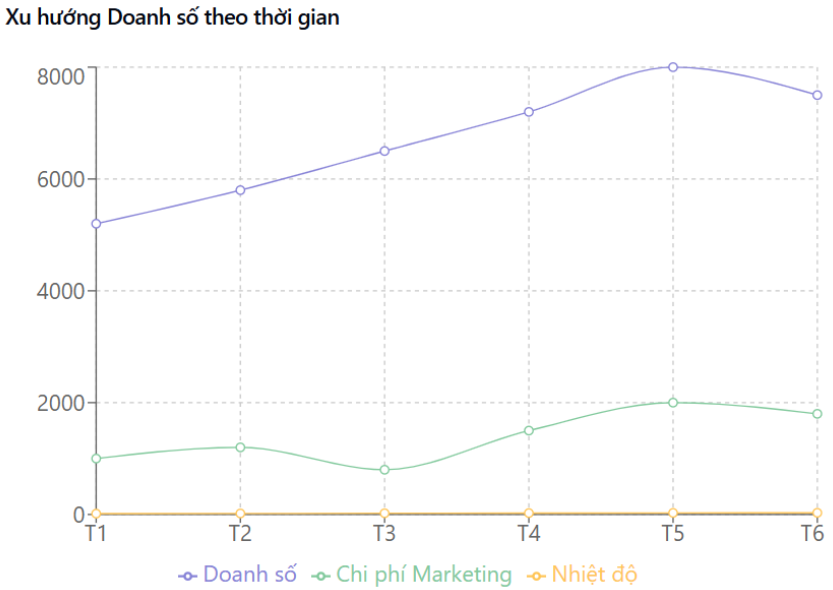
Chẳng hạn như dự đoán giá nhà dựa trên diện tích, dự đoán doanh số bán hàng dựa trên chiến lược tiếp thị.

Số phòng ngủ	Diện tích	Khu đô thị	Giá bán
3	2000	Times City	$250,000
2	800	Royal City	$300,000
2	850	Times City	$150,000
1	550	Times City	$78,000
4	2000	KDT Linh Đàm	$150,000

Mục đích của mô hình

Hiểu rõ mối quan hệ: Xác định mức độ ảnh hưởng của các biến độc lập lên biến phụ thuộc. Điều này giúp hiểu rõ các yếu tố nào là quan trọng và cách chúng ảnh hưởng đến kết quả





Mục đích của mô hình

- ***Đánh giá tác động:*** *Trong nghiên cứu và phân tích, hồi quy tuyến tính giúp đánh giá tác động của các yếu tố (biến độc lập) lên một kết quả cụ thể (biến phụ thuộc).*


2+2=4
A+

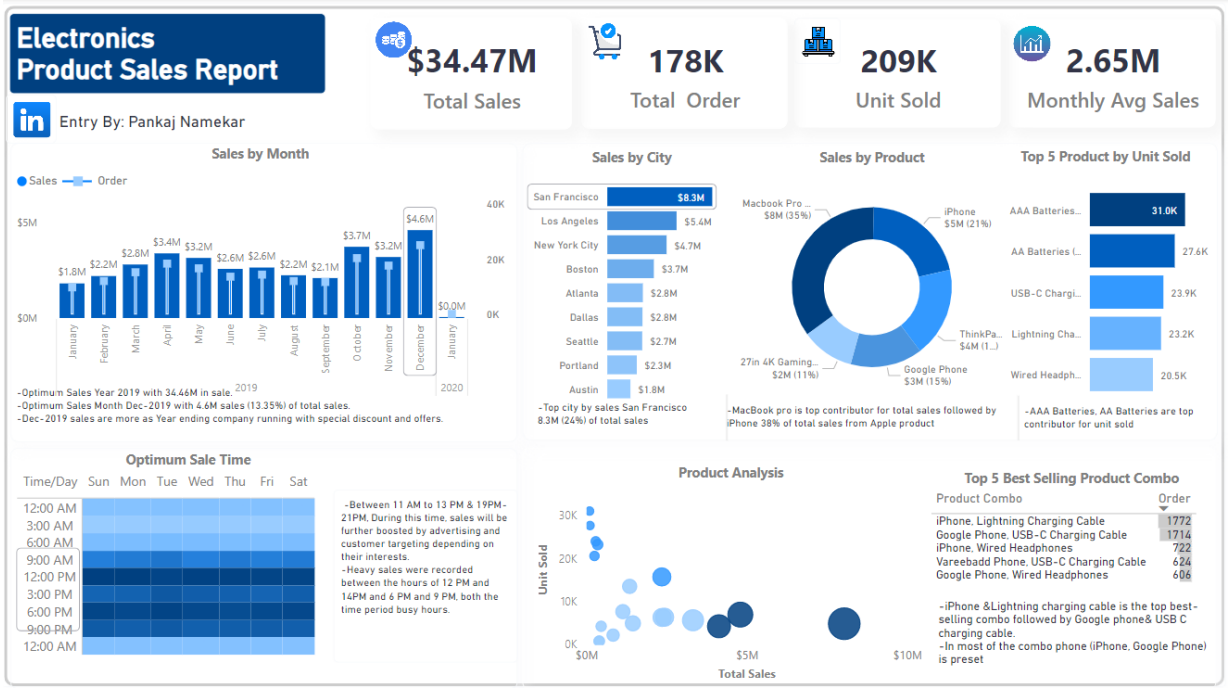

100

Mục đích của mô hình

Kiểm soát các biến: Trong mô hình hồi quy tuyến tính đa biến, có thể kiểm soát ảnh hưởng của các biến nhiễu để tập trung vào mối quan hệ cụ thể giữa một số biến chọn lọc.

Mục đích của mô hình

Phân tích xu hướng và mẫu: Hồi quy tuyến tính giúp phân tích xu hướng và mẫu trong dữ liệu, qua đó cung cấp thông tin hữu ích cho việc ra quyết định và lập kế hoạch.



Giả định cơ bản:

- ***Tính tuyến tính:*** Mối quan hệ giữa biến phụ thuộc và các biến độc lập được giả định là tuyến tính, tức là có thể được mô tả thông qua một đường thẳng.
- ***Độc lập của các sai số (Residuals):*** Các sai số từ mô hình được giả định là độc lập với nhau, không có sự phụ thuộc hoặc mẫu định hình nào.
- ***Phân phối chuẩn của sai số:*** Các sai số được giả định tuân theo một phân phối chuẩn.
- ***Homoscedasticity:*** Phương sai của sai số được giả định là nhất quán qua tất cả các giá trị của biến độc lập, không biến đổi theo mức độ của biến dự đoán.
- ***Không có hoặc hạn chế đa cộng tuyến:*** Giả định rằng không có mối quan hệ tương quan cao giữa các biến độc lập, hay nói cách khác, các biến độc lập không được phụ thuộc lẫn nhau một cách mạnh mẽ.

Tham số

- ***Hệ số hồi quy (regression coefficients)**:* Đây là các trọng số được gán cho mỗi biến độc lập. Chúng xác định mức độ mà mỗi biến độc lập ảnh hưởng đến biến phụ thuộc.
- ***Điểm chặn (intercept)**:* Đây là giá trị của biến phụ thuộc khi tất cả các biến độc lập bằng 0. Nói cách khác, đây là điểm bắt đầu của đường hồi quy trên trục tung.
- ***Lỗi (error term)**:* Còn được gọi là dư lượng, lỗi là phần của dữ liệu mà không được giải thích bởi mô hình hồi quy. Nó bao gồm cả ảnh hưởng của các yếu tố ngoại lai và sai số ngẫu nhiên.

Tham số

- ***Hệ số xác định (R-squared):*** Mặc dù không phải là một tham số cần thiết lập trước khi xây dựng mô hình, R-squared là một chỉ số quan trọng để đánh giá mức độ phù hợp của mô hình với dữ liệu. Nó thể hiện tỷ lệ phần trăm biến thiên của biến phụ thuộc được giải thích bởi mô hình.
- ***Tham số chuẩn hóa (Regularization parameters):*** Đối với các biến thể của hồi quy tuyến tính như Ridge (L2 regularization) hoặc Lasso (L1 regularization), tham số chuẩn hóa được sử dụng để kiểm soát mức độ phạt đối với độ lớn của hệ số hồi quy, nhằm giảm overfitting.
- ***Tiêu chí dừng (stopping criteria):*** Trong các phương pháp học máy, tiêu chí dừng xác định khi nào quá trình tối ưu hóa nên dừng lại, thường dựa trên sự cải thiện của hàm mất mát hoặc số lần lặp tối đa.

Mô hình hồi quy tuyến tính đơn giản (một biến độc lập)

Trong trường hợp đơn giản nhất với chỉ một biến độc lập, mô hình hồi quy tuyến tính có dạng:

$$y = \beta_0 + \beta_1 x + \epsilon$$

trong đó:

- y là biến phụ thuộc.
- x là biến độc lập.
- β_0 là hệ số chặn (intercept).
- β_1 là hệ số hướng (slope).
- ϵ là sai số ngẫu nhiên (không quan sát được).

Mô hình hồi quy tuyến tính đa biến

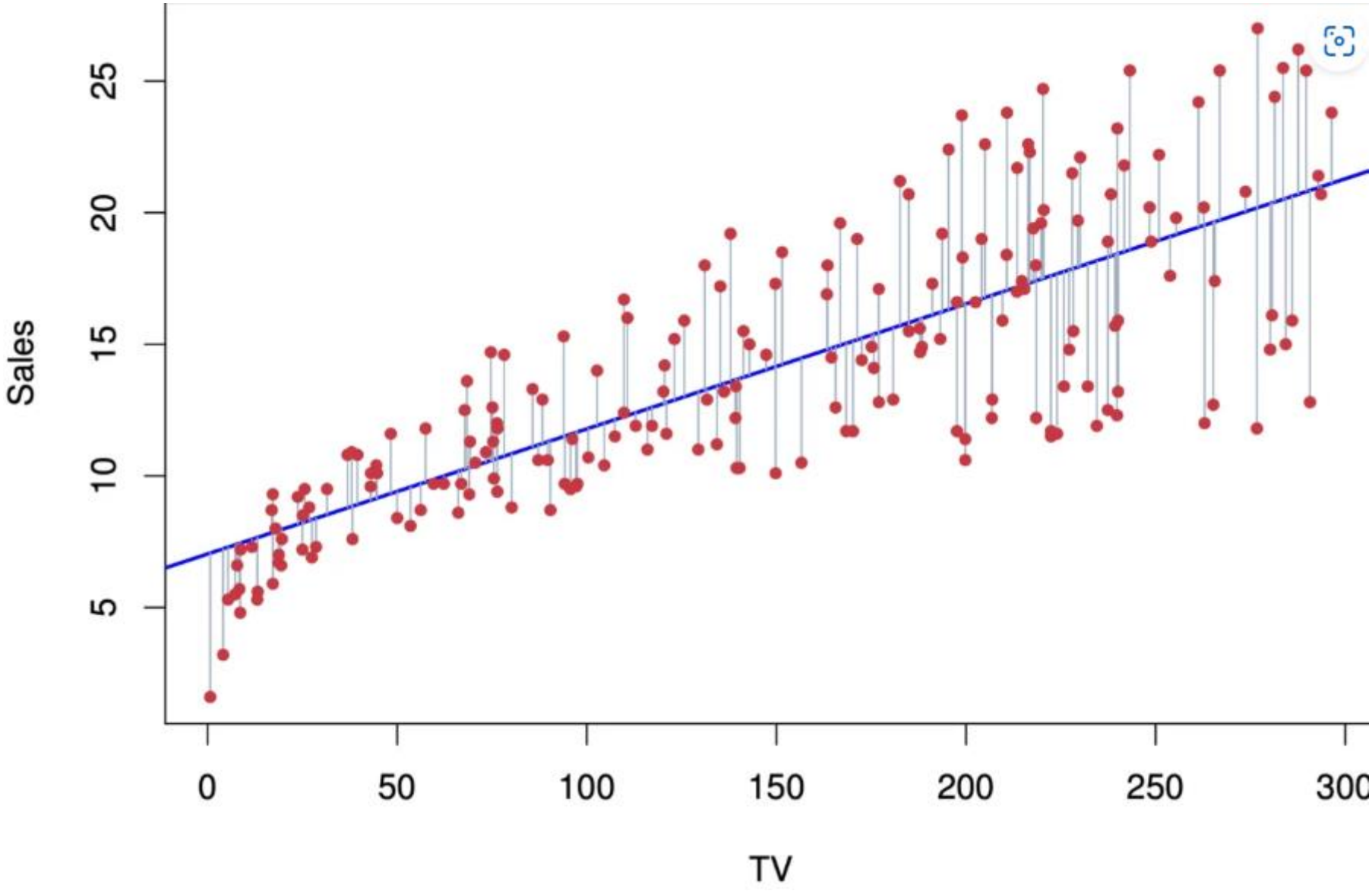
Trong trường hợp có nhiều biến độc lập, mô hình mở rộng thành:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

trong đó, mỗi x_i đại diện cho một biến độc lập khác nhau, và β_i là hệ số tương ứng với mỗi biến độc lập đó.

Tìm hệ số mô hình

Mục tiêu của hồi quy tuyến tính là ***tìm ra các giá trị của hệ số*** β sao cho ***tổng bình phương sai số (sum of squared errors)*** giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất. Phương pháp phổ biến để tìm ra các hệ số này là phương pháp bình phương tối thiểu (least squares method).


Sales
5
10
15
20
25
0
50
100
150
200
250
300
TV

Đánh giá mô hình

Mô hình hồi quy tuyến tính thường được đánh giá dựa trên các chỉ số:

- **R-squared**,
- **Root Mean Squared Error (RMSE)**, hoặc
- **Mean Absolute Error (MAE)**, cho biết mức độ chính xác của mô hình trong việc dự đoán dữ liệu.

Quy trình thực hiện của mô hình

Bước 1:

Thu thập dữ liệu: Bước đầu tiên trong quá trình hồi quy tuyến tính là thu thập dữ liệu. Dữ liệu này có thể đến từ nhiều nguồn khác nhau như khảo sát, ghi chép, cơ sở dữ liệu, v.v. Dữ liệu phải bao gồm cả biến độc lập (predictors) và biến phụ thuộc (target) muốn dự đoán.

linear regression dataset

Notebooks 70,838 | Comments 1,316 | Topics 1,097 | Datasets 911 | Models 29 | Competitions 11

Filter by

DATE

- Last 90 days 4,732
- This week 322
- Today 41

CREATOR

- You 0
- Others 74,202

DATASET SIZE

- small 845
- medium 62
- large 4

DATASET FILE TYPES

74,202 Results Relevance

Simple Linear Regression Algorithm Dataset on Sales Data 123

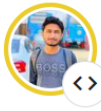
Discussion Topic · 2y ago · by Shruti Pandit

dataset that you can use to apply Simple **Linear** and Advanced **Linear Regression** Techniques and check underlying 15 comments

Linear Regression - Salary Dataset 102

Notebook · 3y ago · by Shubham

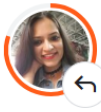
/input/salary-data-**dataset**-for-**linear**-**regression**/Salary_Data.csv") **EDA** df.head() #check number of 7 comments



Tip Prediction on Tips dataset (Linear Regression) 108

Notebook · 9mo ago · by Atif Ali AK

using **Linear Regression** - **Task** : - To predict the tip amount on the basis of total bill, sex 34 comments



Comment on: Simple Linear Regression Algorithm Dataset on Sales Data 0

Discussion Comment · 2y ago · by Shruti Pandit

In the General forum 1 reply

Stacked Regressions : Top 4% on LeaderBoard 7364

Notebook · 7y ago · by Serigne

[A study on **Regression** applied to the Ames **dataset**][2] by **Julien Cohen-Solal** : Thorough features 1,134 comments

Quy trình thực hiện của mô hình

Bước 2:

Chuẩn bị dữ liệu: Sau khi thu thập, dữ liệu cần được làm sạch và chuẩn bị. Điều này bao gồm việc loại bỏ hoặc xử lý dữ liệu thiếu hoặc nhiễu, chuẩn hóa hoặc tiêu chuẩn hóa các biến độc lập, và chuyển đổi dữ liệu (nếu cần).

Quy trình thực hiện của mô hình

Bước 3:

Huấn luyện mô hình: Dựa trên dữ liệu huấn luyện, mô hình hồi quy tuyến tính được xây dựng bằng cách ước lượng các hệ số cho mỗi biến độc lập. Quá trình này thường bao gồm việc sử dụng phương pháp bình phương nhỏ nhất (least squares method) để tìm ra đường thẳng (hoặc mặt phẳng, siêu phẳng) phù hợp nhất với dữ liệu.

Quy trình thực hiện của mô hình

Bước 4:

Đánh giá mô hình: Mô hình được đánh giá dựa trên hiệu suất của nó trên dữ liệu kiểm thử. Các chỉ số đánh giá thường dùng bao gồm R-squared (đo lường mức độ "phù hợp" của mô hình với dữ liệu), Mean Squared Error (MSE), hoặc Root Mean Squared Error (RMSE).

Quy trình thực hiện của mô hình

Bước 5:

Tinh chỉnh mô hình: Dựa trên kết quả đánh giá, mô hình có thể cần được tinh chỉnh để cải thiện hiệu suất. Điều này có thể bao gồm việc điều chỉnh các biến đầu vào, sử dụng các kỹ thuật regularization (như Ridge hoặc Lasso), hoặc thử nghiệm các mô hình hồi quy khác nhau.

Quy trình thực hiện của mô hình

Bước 6:

Sử dụng mô hình: Cuối cùng, mô hình được sử dụng để thực hiện dự đoán trên dữ liệu mới. Kết quả của quá trình này cho phép ta ứng dụng những phát hiện từ mô hình hồi quy vào thực tế, dự đoán kết quả hoặc hiểu rõ hơn về mối quan hệ giữa các biến. Mỗi bước trong quá trình này đều quan trọng và cần được thực hiện cẩn thận để đảm bảo tính chính xác và hiệu quả của mô hình hồi quy tuyến tính.

Ví dụ

Giả sử muốn dự đoán giá nhà dựa trên diện tích (m²) của nó. Trong trường hợp này, giá nhà là biến phụ thuộc (y), và diện tích nhà là biến độc lập (x).

Bước 1 (Thu thập dữ liệu): Thu thập dữ liệu về các ngôi nhà bao gồm giá bán và diện tích của chúng như sau:

Diện tích (m²)	Giá (nghìn USD)
50	200
70	270
80	300
100	370
120	450

Ví dụ

Bước 2 (Chuẩn bị dữ liệu): Trong trường hợp này, dữ liệu đã khá sạch và không cần xử lý nhiều.

Bước 3 (Phân chia dữ liệu): Sử dụng 80% dữ liệu để huấn luyện mô hình và 20% còn lại để kiểm thử mô hình.

Bước 4 (Huấn luyện mô hình): Sử dụng phương pháp bình phương nhỏ nhất để tìm ra hệ số cho mô hình hồi quy tuyến tính. Giả sử mô hình tìm được là:

Giá = 50+3×Diện tích

Bước 5 (Đánh giá mô hình): Kiểm tra mô hình với 20% dữ liệu còn lại và tính toán các chỉ số như R-squared, MSE để đánh giá hiệu suất.

Bước 6 (Tinh chỉnh mô hình): Điều chỉnh mô hình dựa trên kết quả đánh giá.

Bước 7 (Sử dụng mô hình): Sử dụng mô hình để dự đoán giá của nhà dựa trên diện tích. Chẳng hạn, với một ngôi nhà có diện tích 85 m^2, mô hình sẽ dự đoán giá là: 50 + 3 × 85 = 305 (nghìn USD).

Bài tập

- Thực hành với bộ dữ liệu giả định

- Phân tích dữ liệu bằng mô hình hồi quy tuyến tính trên bộ dữ liệu thực IRIS

- Phân tích dữ liệu bằng mô hình hồi quy tuyến tính trên bộ dữ liệu giả định MTCARS