Artificial Intelligence

Linear Regression vs. Logistic Regression: Understanding 13 Key Differences

Linear regression is utilized for regression tasks, while logistic regression helps accomplish classification tasks.

Vijay Kanade AI Researcher

Last Updated: June 10, 2022

> ***Supervised machine learning is a widely used machine learning technique that predicts future outcomes or events. It uses labeled datasets to learn and generate accurate predictions. Supervised learning is classified into two categories, namely, regression and classification.***

Regression is a model that predicts continuous values (numerical), while classification mainly classifies the data. Regression is accomplished by using a linear regression algorithm, and classification is achieved through logistic regression. This article highlights the critical differences between linear and logistic regression.

Table of Contents

Linear Regression vs. Logistic Regression

Supervised machine learning is a widely used machine learning technique that predicts future outcomes or events. It uses labeled datasets to learn and generate accurate predictions. Supervised learning is classified into two categories, namely, regression and classification.

Regression is a model that predicts continuous values (numerical), while classification mainly classifies the data. Regression is accomplished by using a linear regression algorithm, and classification is achieved through logistic regression.

Linear regression

A linear regression algorithm defines a linear relationship between independent and dependent variables. It uses a linear equation to identify the line of best fit (straight line) for a problem, thereby enabling the visualization and prediction of the output of the dependent variables.

For example, linear regression may predict how the obesity of an individual is linearly related to work-life imbalance.

Linear regression is further subdivided into simple and multiple linear regression, wherein a single and two or more independent variables are respectively used to predict the output.

Logistic regression

Logistic regression's output lies between 0 and 1 as the algorithm is designed to predict a binary outcome for an event based on the previous observations of a data set. It uses independent variables to predict the occurrence or failure of specific events.

For example, logistic regression predicts whether a patient has stage 2 (0) or stage 3 (1) cancer.

See More: What Is Artificial Intelligence (AI) as a Service? Definition, Architecture, and Trends

Understanding the 13 Key Differences Between Linear and Logistic Regression

Linear and logistic regression are extensively used to accomplish data science tasks; however, each model addresses specific problems. Unlike the linear model, logistic regression uses a complex equation model that makes it harder to understand and interpret. Apart from the equation model, linear and logistic regression differ significantly.

Let’s understand the key differences between the linear and logistic regression models.

1. Variable & output type

A linear regression model relies on a *continuous dependent variable*. This implies that the dependent variable takes up numeric values instead of being classified under categories or groups. In contrast, logistic regression models rely on binary dependent variables. The dependent (or response) variable can take up only two values – 0 or 1.

Also, linear regression output has a continuous value (it gives a range of values). For example,

- Length of the roof (25 inches, 19 inches, 5 ft)
- Height (5 ft 8 inches, 6 ft 2 inches, 5 ft 10 inches)
- Escape velocity (26000 mph, 21500 mph, 29500 mph)

On the other hand, the logistic regression model is revealed via probabilities. For example,

- 84.3% chance of losing a tennis match
- 23.1% chance of passing a bill in Congress
- 65.1% chance of imposing a curfew during a COVID-19 outbreak

Moreover, linear regression observes a *normal or gaussian distribution*, and logistic regression reveals a *binomial distribution*.

2. Relationship between variables

Understanding the relationship between variables is crucial when deciding the type of regression model to be used for different purposes.

Linear regression describes a linear relationship between variables by plotting a straight line on a graph. It enables professionals to check on these linear relationships and track their movement over a period. On the contrary, logistic regression is known to study and examine the probability of an event occurrence. Since it does not denote a linear structure of a variable relationship, tracking logistic regression using linear structures is not required.

3. Mathematical equation

The linear relationship between variables (i.e. predictor and response) for linear regression models can be interpreted by the following equation:

$y = a0 + a1x1 + a2x2 + \dots + aixi$

Here,

- y denotes response variable
- xi denotes ith predictor variable
- ai denotes the average effect on y as xi increases by one (keeping all the predictors fixed)

Similarly, logistic regression predicts the probability events or observation through the following equation:

$y(x) = e(a0 + a1x1 + a2x2 + \dots + aixi) / (1 + e(a0 + a1x1 + a2x2 + \dots + aixi))$

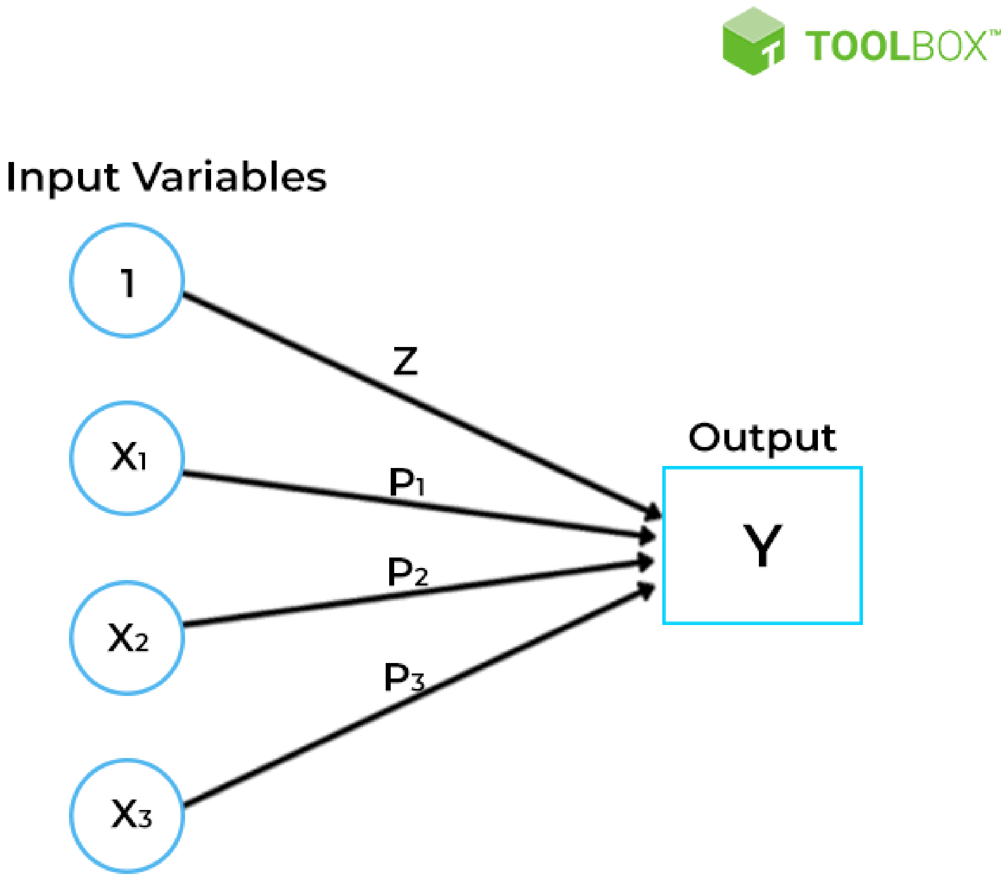
4. Methods employed to fit equation

A linear regression model uses an ‘*ordinary least squares*’ method to determine the best fitting regression equation. As per the method, the regression coefficients should be chosen to lower the sum of the squared distances of every response variable to the fitted value.

On the contrary, logistic regression uses the ‘*maximum likelihood estimation*’ method, where the regression coefficients are chosen to maximize the probability of y for a given x (likelihood). In the context of ML, the system performs several iterations until the maximum likelihood estimates are achieved.

5. Kind of predictions

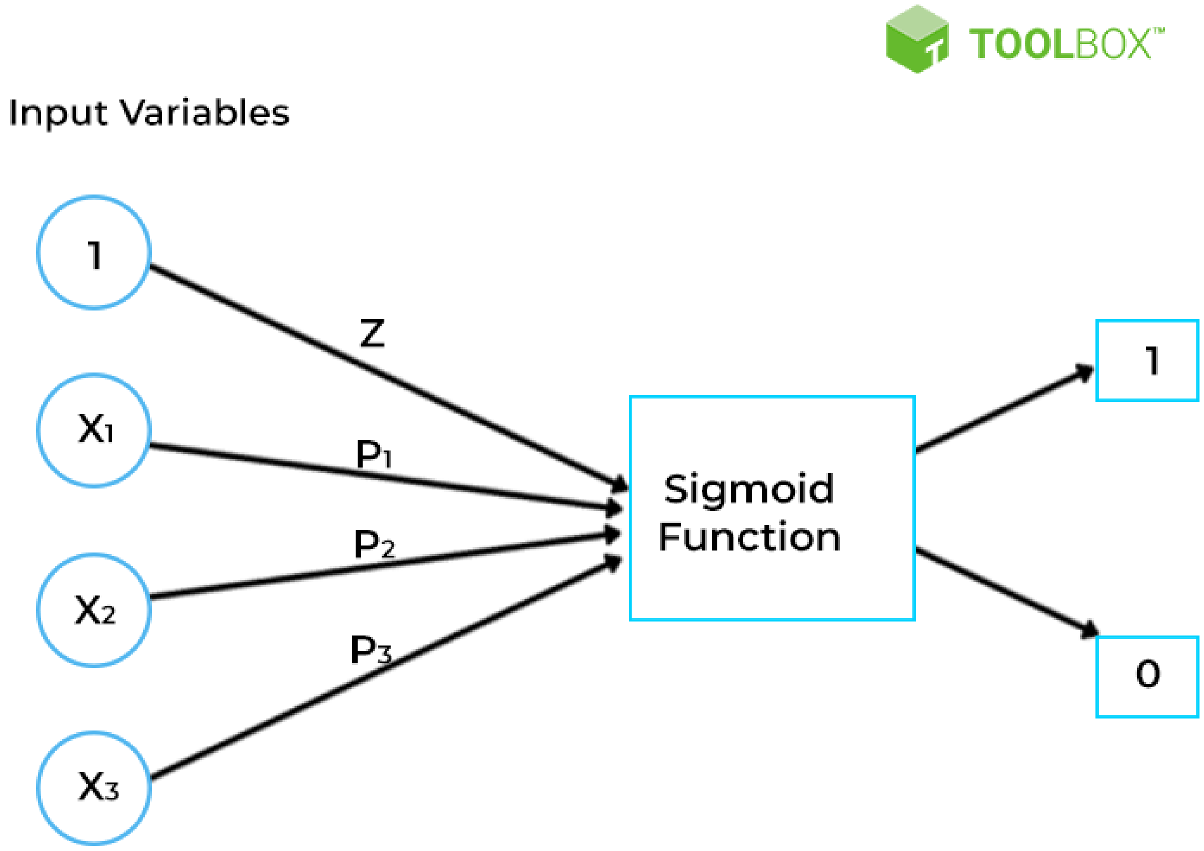
A linear regression model estimates an output ‘y’ (real value) by considering the sum of the values of all input features (variables).



Output (y) = z + p1x1 + p2x2 + p3x3 +……..+ pnxn

The model determines the values for coefficients z, p1, p2, p3….pn and subsequently fits the training data to predict the real-valued output (y) with minimal error.

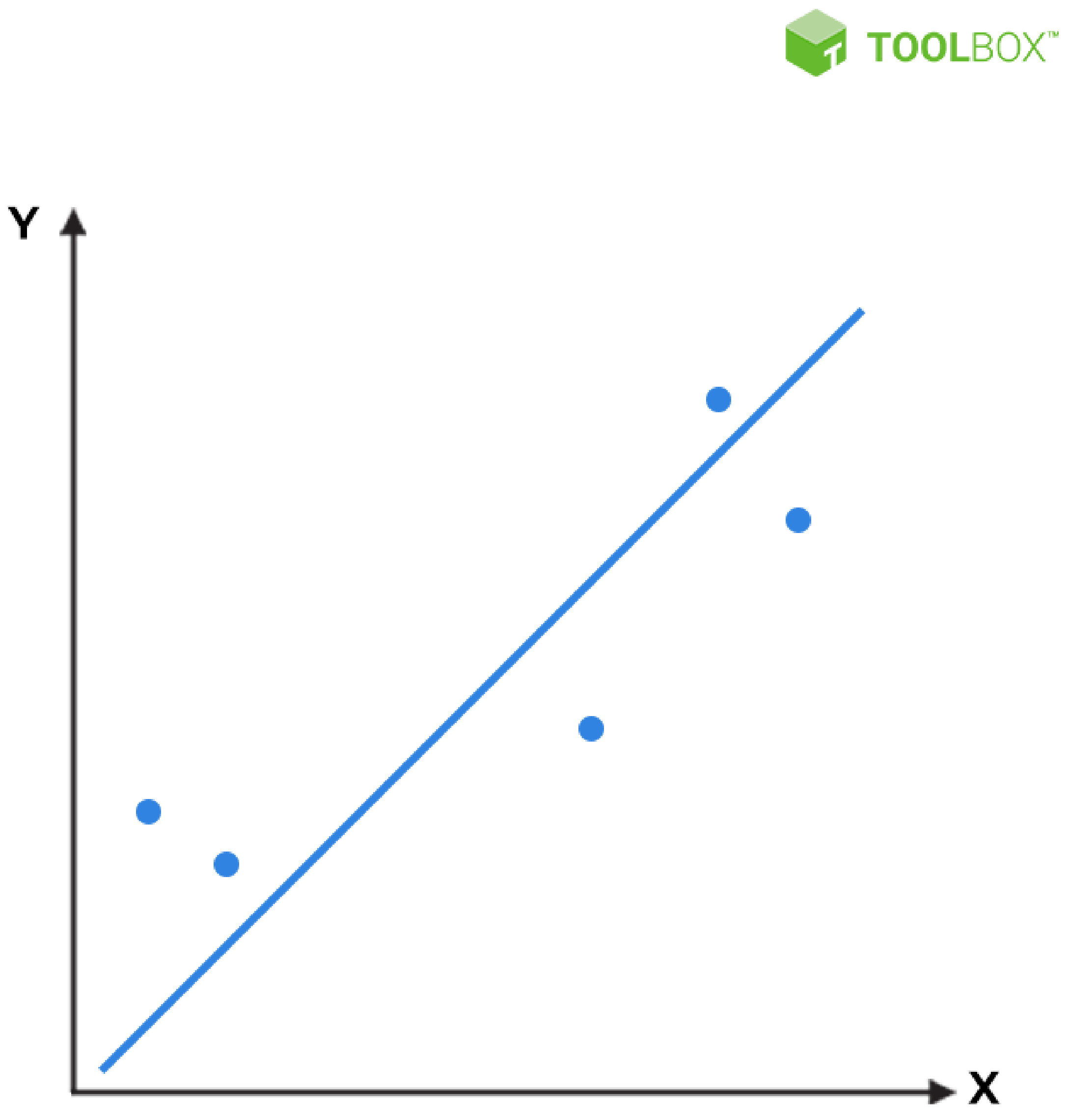
Conversely, a logistic regression model considers the sum of the input variables' values and applies a logistic function or *sigmoid function* to the result. The non-linear function thereby yields a binary output in the form of 0 or 1 (or even 'true or false').



y = logistic (z + p1x1 + p2x2 + p3x3 +…….+ pnxn)
y=1/1+e^[−(z + p1x1 + p2x2 + p3x3 +…….+ pnxn)]

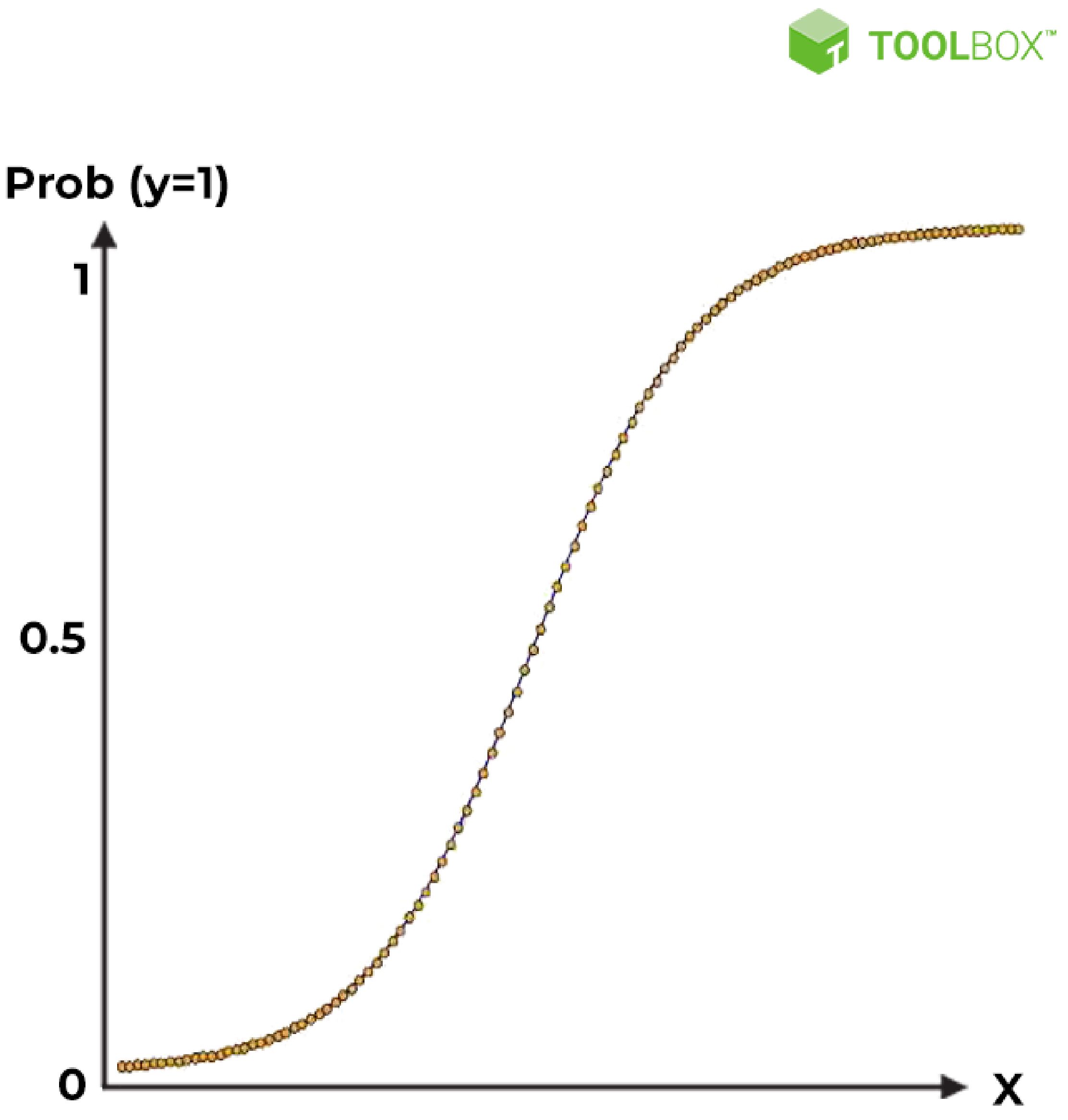
6. Graphical representation: Curve

Linear regression is represented by a straight line, also termed a *regression line*. This line reveals the predicted score on 'y' for each 'x' value. Moreover, the distance of the data points on the plot from the regression line discloses errors in the model.



Graphical Representation of Linear Regression

Conversely, logistic regression reveals an *S-shaped curve*. Here, change in regression coefficients has an impact on the curve direction and its steepness. Thus, one can infer that a positive slope results in an S-shaped curve, and a negative slope reveals a Z-shaped curve.



Graphical Representation of Logistic Regression

7. Correlation between independent variables

A linear regression testing model effectively determines the correlations between multiple variables. A simple linear regression tends to define the correlation between dependent and independent variables. Also, in multiple linear regression, you can identify one or more possible correlations between variables. The correlation phenomenon is much like the cause-and-effect relationship.

On the other hand, in logistic regression, independent variables do not share any correlations. Here, the independent variables are all independent and lack any dependent variables for any correlation to occur.

8. Weights of next observations

Linear regression uses the root-mean-square error (RMSE) to calculate the next weight value of the data points (or observations) spread across the regression line. Conversely, logistic regression uses a precision method to predict the next weight value. The RSME method effectively evaluates the accuracy of the linear model and helps determine the prediction errors shown by the logistic model.

9. Activation functions

Regression models in machine learning use different *activation functions* to signal an artificial intelligence network to activate a specific neuron. A linear regression model does not require any such activation function. However, while converting a linear model into a logistic model, an activation function becomes essential.

In logistic regression, the sigmoid function is used as an activation function, converting the outcome into a categorical value. The function activates the system or AI network when specific parameters or criteria are met.

10. Interpretability

The linear and logistic probability models are given by the following equations:

p = a0 + a1x1 + a2x2 + … + aixi ———(1) [linear model]

ln[p/(1-p)] = b0 + b1x1 + b2x2 + … + bkxk —— (2) [logistic model]

Where p = probability.

From equations 1 and 2, you can say that probability (p) is considered a linear function of the regressors for the linear model. Whereas, for the logistic model, the log odds p/(1-p) are considered a regressors' linear function.

By observing the above equations, one can say that the linear model is more interpretable than the logistic model. For example, consider a1 as 0.07 in equation 1. This implies that a single unit increase in x1 causes a 7 percentage point increase in the probability of y.

Now consider equation 2 of the logistic model. Here, if we consider b1 as 0.07, this implies a 0.07 increase in the log odds of y for a single unit increase in x1. This makes it complex to interpret the overall scenario.

11. Rule of thumb

In situations when you are modeling extreme probabilities–probability closer to 0 or 1–you can prefer logistic regression. However, if the probabilities have intermediate values, say between 0.30 and 0.70, you can opt for linear regression. Even though both linear and logistic regression perform equally well in this case, linear regression is more straightforward to interpret than the logistic model.

Consider a use case where you're modeling the probability of a survey. Almost all the modeled probabilities will inevitably lie between 0.25 and 0.75. A linear probability model may be suitable here due to straightforward interpretation.

Conversely, if you consider modeling the probability of an ATM transaction being fraudulent or not, the modeled probabilities will lie between 0.000003 and 0.25. This use case is tailor-made for the logistic model, with the linear model not performing well here.

12. Computational speed

Both linear and logistic models operate at different computing speeds. Logistic regression uses an iterative process of maximum likelihood to fit the model, making it slower from the outset. This slowness in computing speed might not be evident when using a small-sized dataset or fitting a simple model. However, the situation worsens when a larger dataset comes into the picture or while fitting a complex model.

On the contrary, the linear probability model is faster than the logistic model as it can be predicted non-iteratively by employing the 'ordinary least squares' method.

13. Applications

The two regression models are used in a diverse set of applications. In specific, linear regression finds application in data science, business, finance, and marketing.

- **Business insights**: Companies employ linear regression to develop *business insights* that help them streamline and optimize their operations according to market trends. The various parameters tracked down via linear regression include evaluating trends, determining consumer inclination or behavior, performing sales forecasts, and estimating profit or loss margins. In a way, these linear models boost the overall performance of businesses.
- **Market analysis**: A comprehensive market analysis is performed by businesses with the help of linear regression models. Various marketing strategies are evaluated as companies zero in on factors that affect the overall sales of a product or service, such as product pricing, design, marketing campaigns, and promotions.
- **Financial risk assessment**: Linear regression is widely used by analysts in the financial industry vertical. It is used for forecasting returns, portfolio management, and asset valuation. These models play a crucial role in determining the relationship between assets' estimated returns and the associated market risk.

Similarly, logistic regression has the following applications:

- **Medicine**: Logistic regression can be used by medical practitioners who intend to study the effect of jogging and intense running on the probability that an athlete may endure a knee injury. The response variable, in this case, will equate to 'knee injury' and give two possible results:

1. A severe knee injury.
2. A mild knee injury.

The model's outcome will justify how jogging and intense running can affect the knee injury probability for an athlete.

- **Credit scoring**: Automated credit scoring is achievable by developing predictive models through logistic regression. Various parameters such as account status, credit history, marital status, gender, and others are considered while calculating credit scores. With the increase in the number of variables considered, logistic models are bound to give accurate predictions in most cases.

- **Customer behavior tracking:** Logistic regression is a machine learning model that tracks customer behavior across online platforms, social media services, and even video gaming services. These models are fast, self-learning, and easy to interpret. It makes them ideal for automating functions and improving the overall end-user experience.

- **Hotel bookings**: Logistic regression is used by several hotel booking sites and applications to predict travelers' behavior, interests, and intentions. This data is used to recommend future holiday destinations and accommodation choices for travelers. The models rely on users' historical data that reveals how they interact with their sites. This boosts the chances of users finalizing a particular hotel for their next holiday season.

- **Text editing**: Logistic regression models are used by several text editing tools that identify and correct errors of different types, such as grammatical errors, syntactical errors, typos, and even structural mistakes in sentences. One can also train these models to spot and flag offensive words, morally-insensitive words, and others based on the user's choice and available software features.

See More: Top 10 AI Companies in 2022

Takeaways

Regression analysis identifies the trends in data by revealing a relationship between a set of dependent and independent variables. It is a crucial machine learning and statistical analysis tool that predicts outcomes, forecasts data, and determines the dependencies between variables.

Both linear and logistic regression represent the two types of this very regression analysis, where linear regression predicts a continuous outcome while logistic regression yields a discrete value. In simple words, regression is accomplished with linear regression, while classification is achieved through logistic regression.