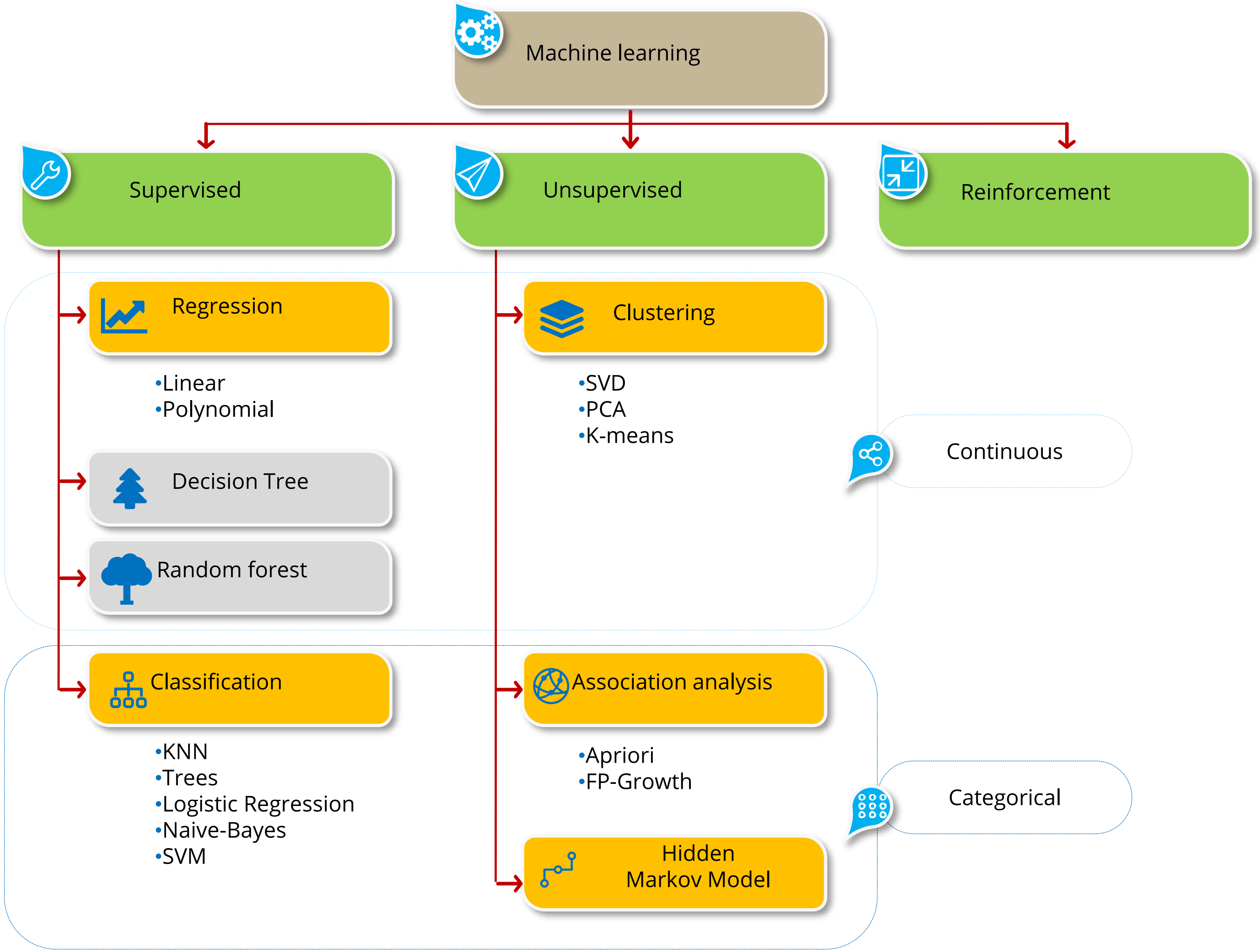
Deloitte.

Citizen Data Scientist – Unit 2 | Part 1
Supervised Learning

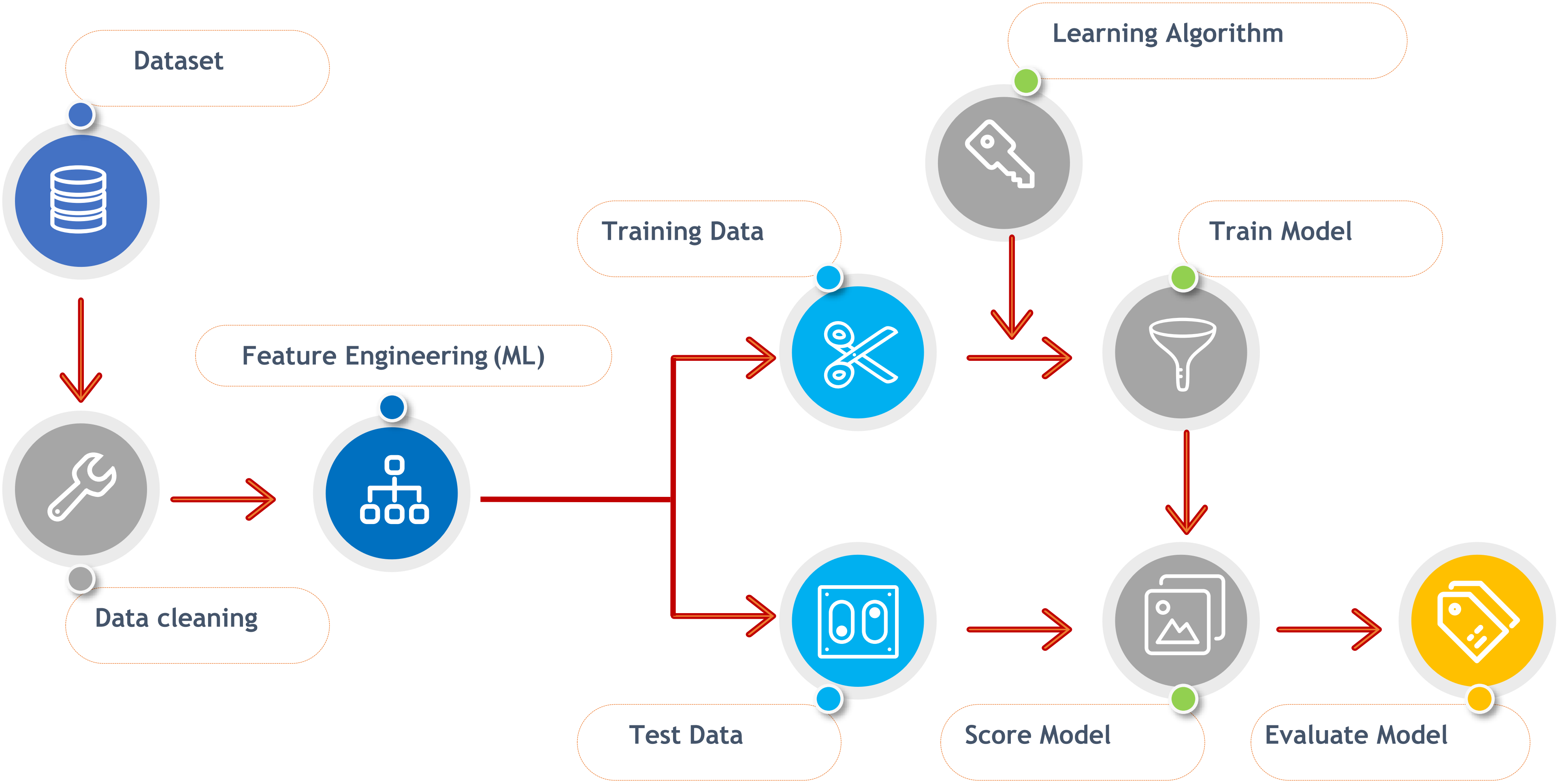
April 2024

MAKING AN
IMPACT THAT
MATTERS
since 1845

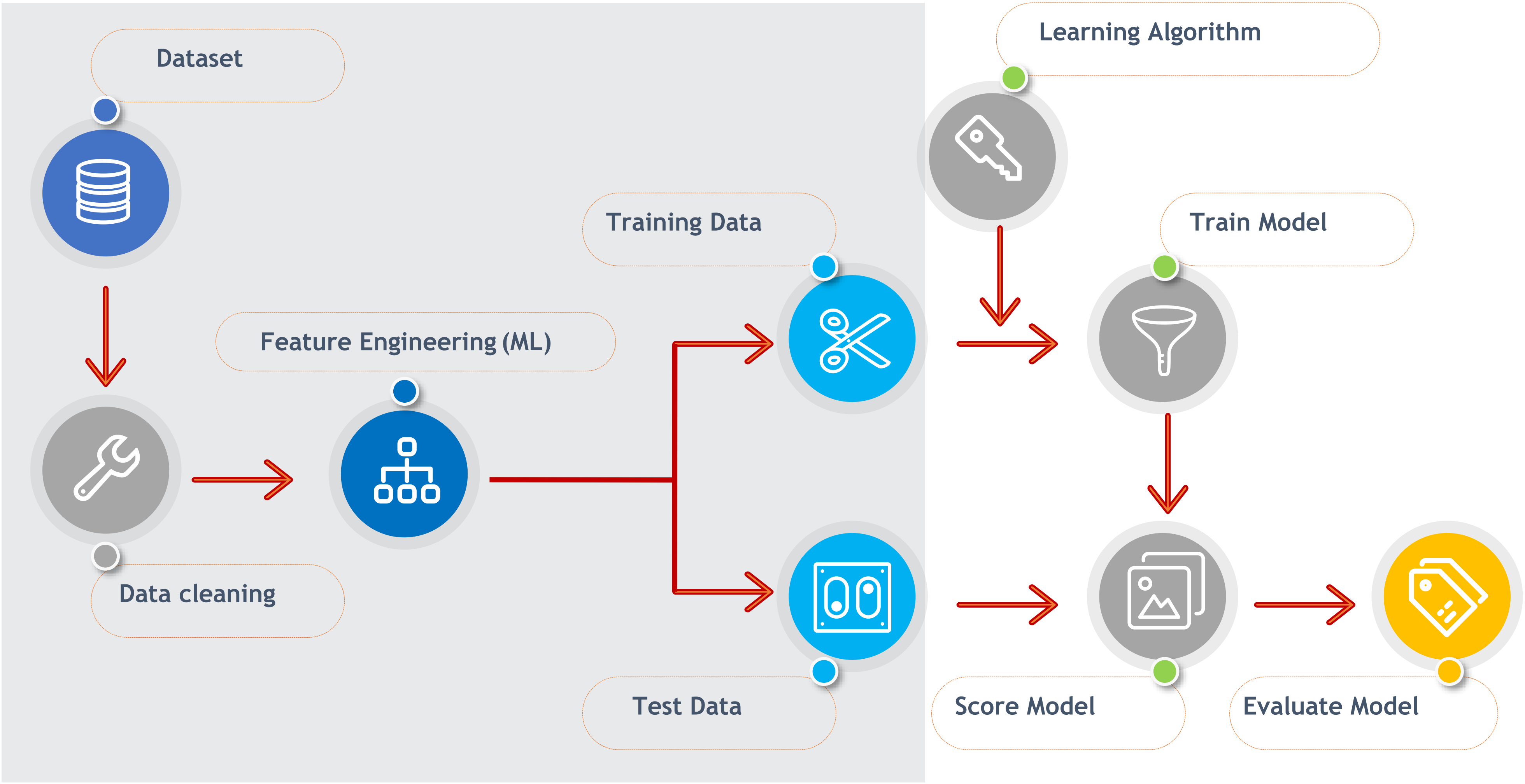
Types of Machine Learning



Supervised Learning Process



Supervised Learning Process



The Importance of Splitting Data into Training and Test Sets

The Importance of Splitting Data into Training and Test Sets

In the world of Supervised Learning , the division of data into training and test sets is a critical step that must not be overlooked. This process is the backbone of developing a model that is not only accurate but also generalizable to new data — a fundamental goal in predictive modeling.

Why Split Data?

- **Model Assessment:** Splitting data into separate training and testing sets allows us to train our models on one subset of data and then evaluate their performance on another, untouched subset. This gives us insight into how the model will perform in the real world.
- **Overfitting Prevention:** Overfitting occurs when a model learns the details and noise in the training data to the extent that it negatively impacts the performance of the model on new data. The test set acts as a check against this, ensuring that the model's predictions are actually due to learning and not memorization.
- **Model Tuning:** The test set provides a final, unbiased performance metric for model selection. This is crucial when tuning hyperparameters or making decisions about which features to include in your model.

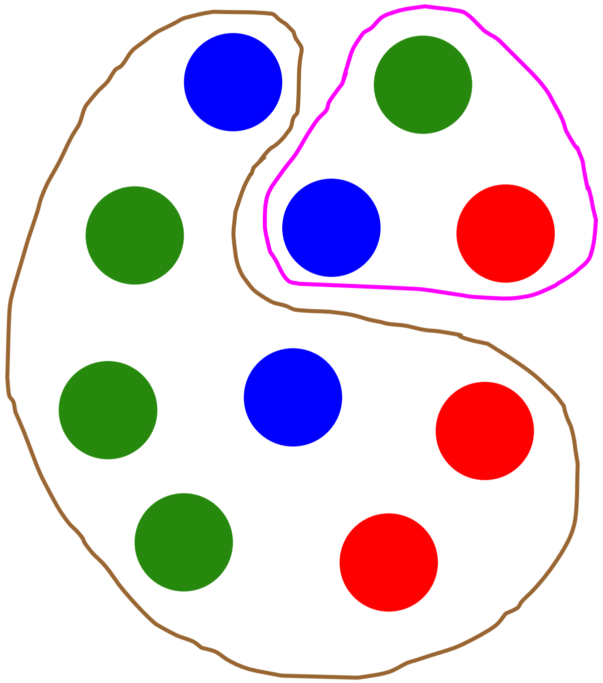
Best Practices for Splitting Data

- **Random Sampling:** Data should be randomly divided to prevent any biases that could skew the results of the test set performance. Most machine learning libraries provide functions to shuffle data before splitting.
- **Stratified Sampling:** When dealing with classifications, particularly with imbalanced classes, stratified sampling ensures that the train and test sets have the same proportion of class labels as the original dataset.
- **Temporal Considerations:** For time-series data, random splitting is not ideal. Instead, you should use chronological splits to maintain the temporal sequence of observations.
- **Size Matters:** Typically, the training set is larger than the test set. A common split ratio is 80:20 or 70:30, training to test, although this can vary based on the size of the dataset.
- **Cross-Validation:** While not a method of splitting, cross-validation involves dividing the data into multiple blocks and rotating which block is used for testing vs. training. This method helps maximize the use of available data.

Splitting the Data

Training Set

The training set is utilized to train the model. This subset of data allows the model to learn and adapt its parameters for optimal performance. The quality and comprehensiveness of the training set are crucial, as they directly influence how well the model can capture the underlying patterns in the data.



Test Set

The test set is used to evaluate the model after the training process has concluded. It acts as new, unseen data, allowing us to simulate how the model will perform in real-world scenarios. This set helps verify the model's effectiveness and generalizability, ensuring that its predictions are robust and reliable.

Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10

Regression

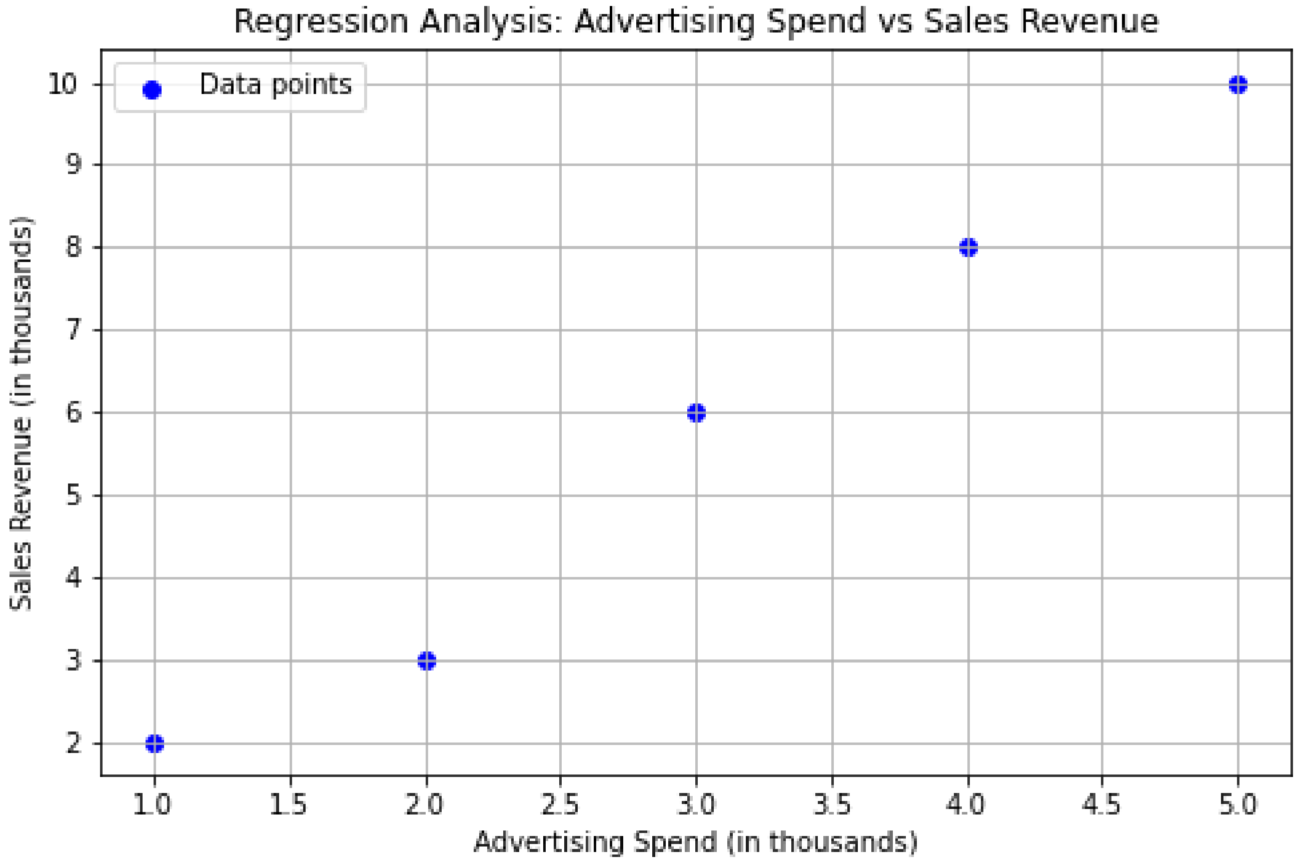
Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10

Sales Revenue = target = y

Advertising Spend = predictor = X

Is it possible to estimate a mathematical model?



Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10

Sales Revenue = target = y

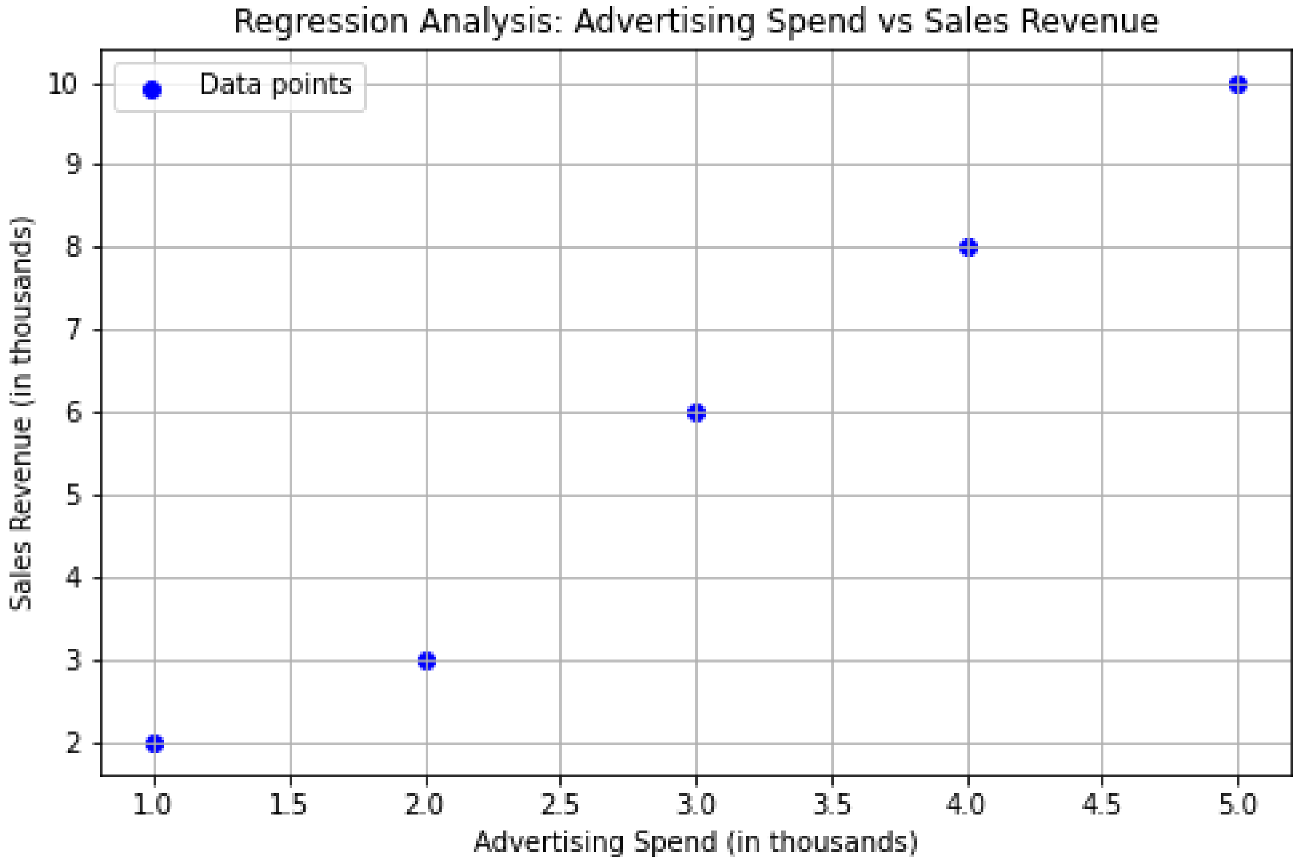
Advertising Spend = predictor = X

Is it possible to estimate a mathematical model?

y = mX + b

m = angular coefficient (slope)

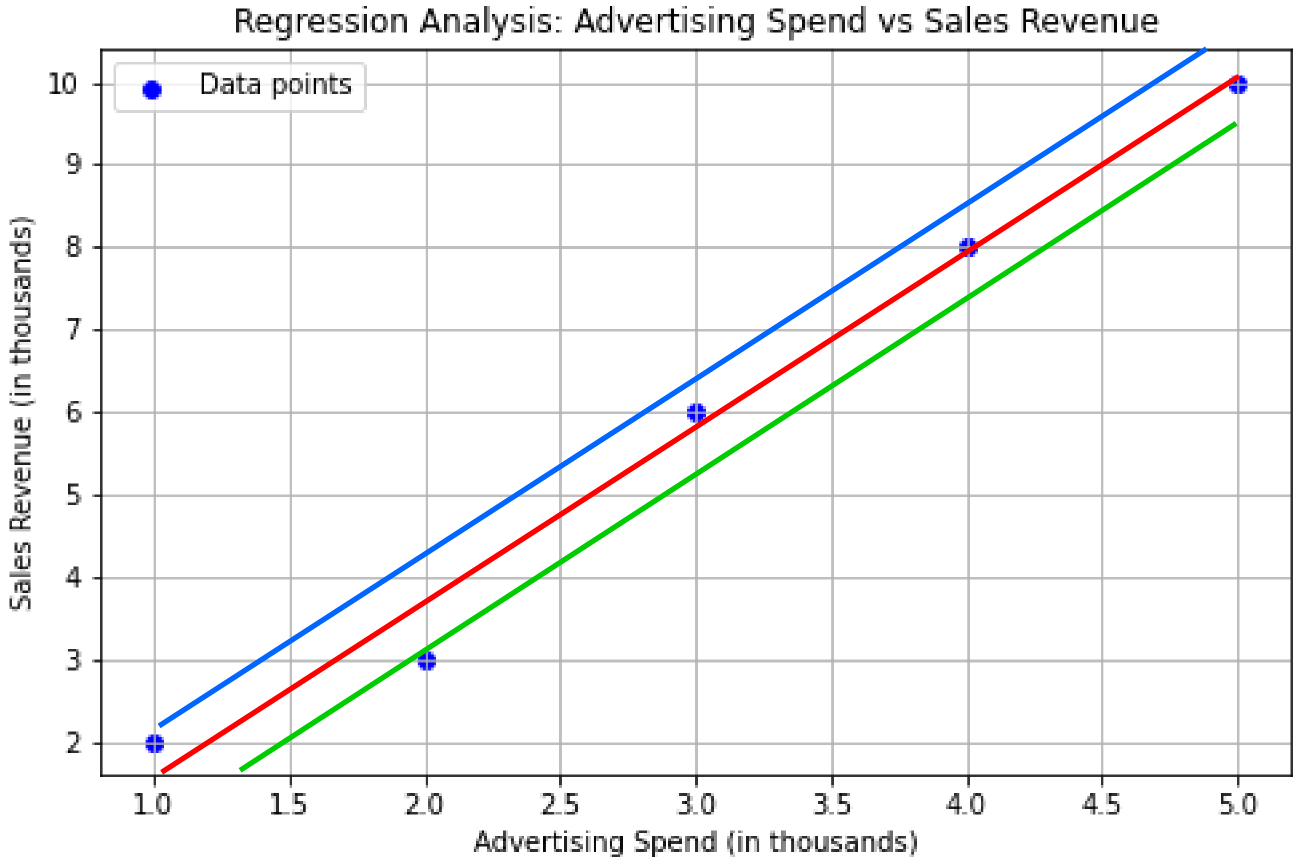
b = where the line intercepts y axis



Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10

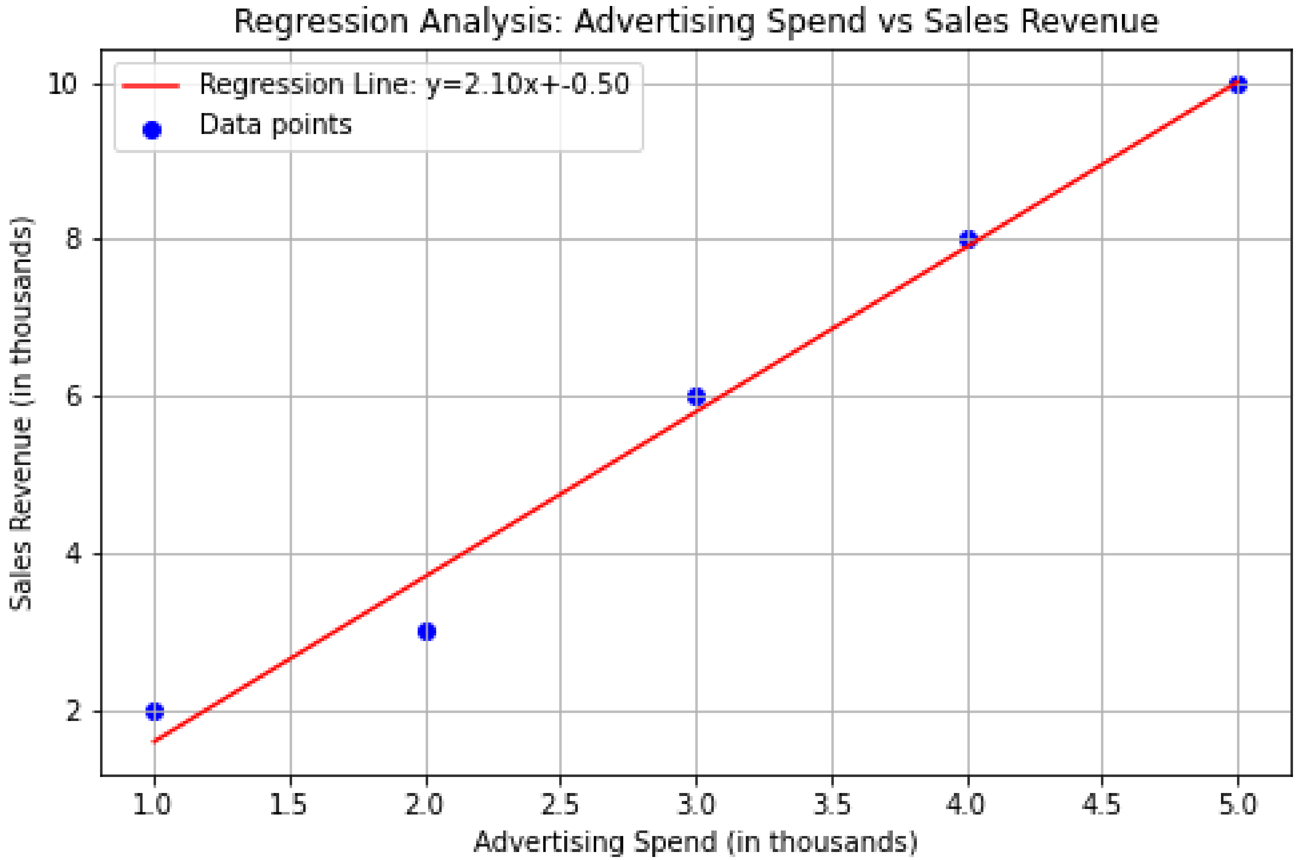


Which line is better?

Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

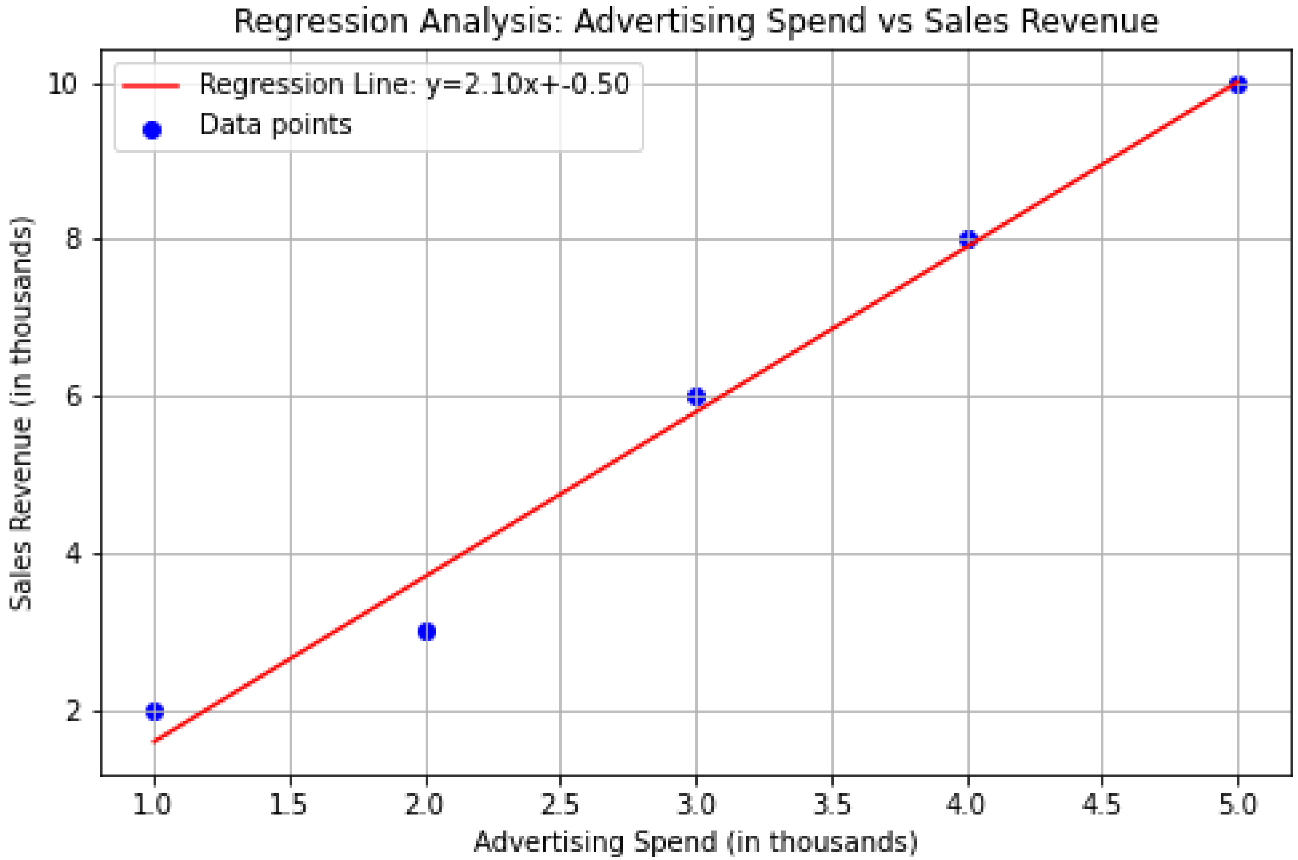
Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10



Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10
7	**?**



Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

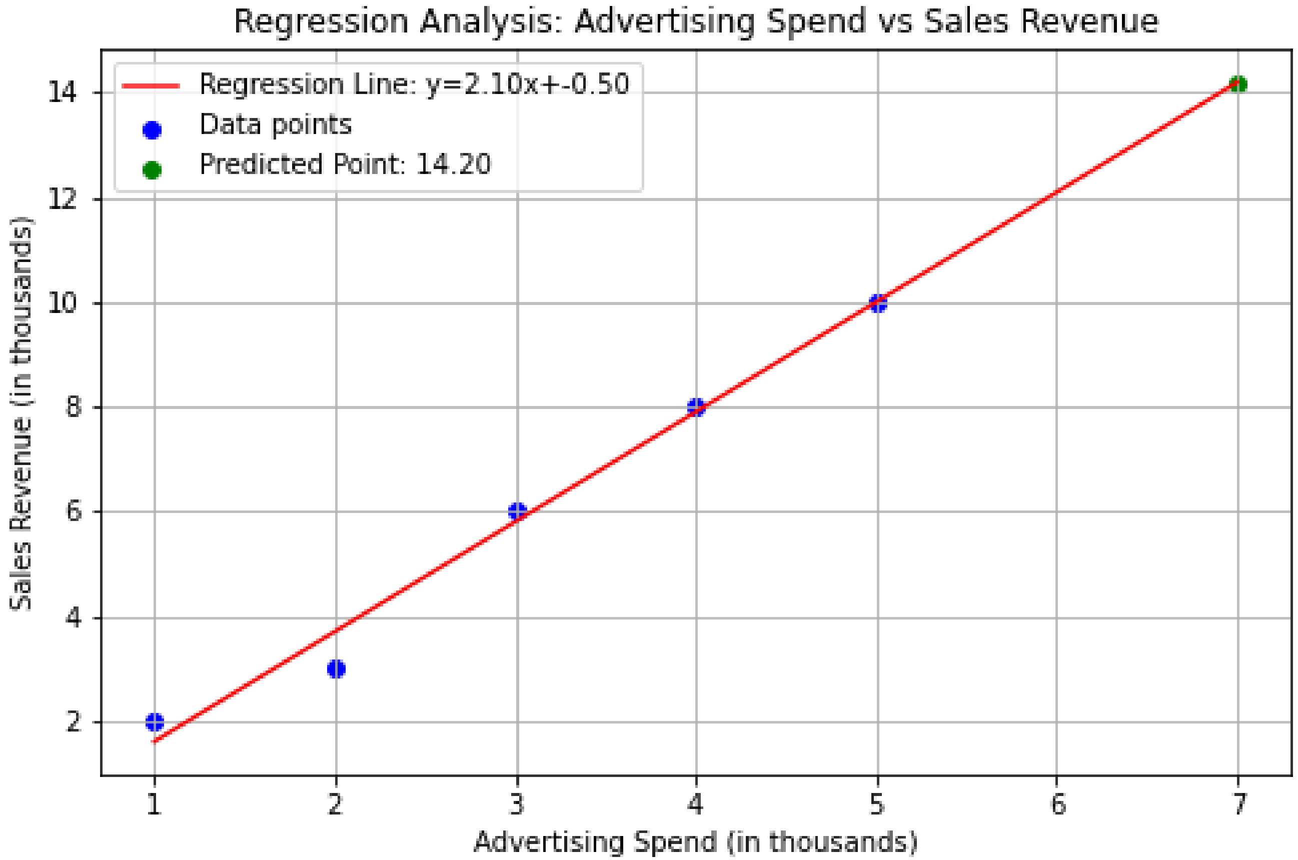
Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10
7	**14.20**

y = mX + b

m = 2.10 and b = -0.50

y_pred = 2.10 * 7 - 0.50

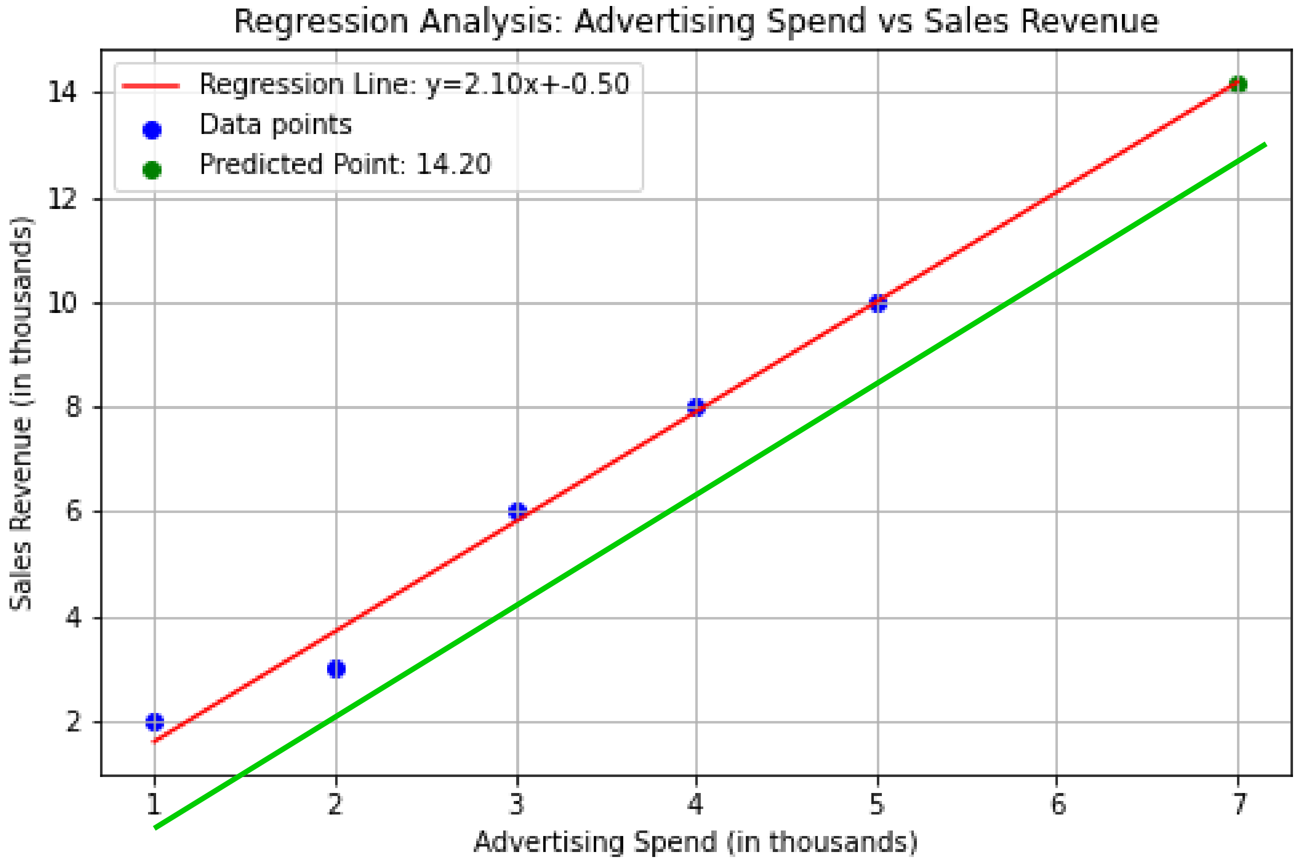
y_pred = 14.20



Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10
7	14.20

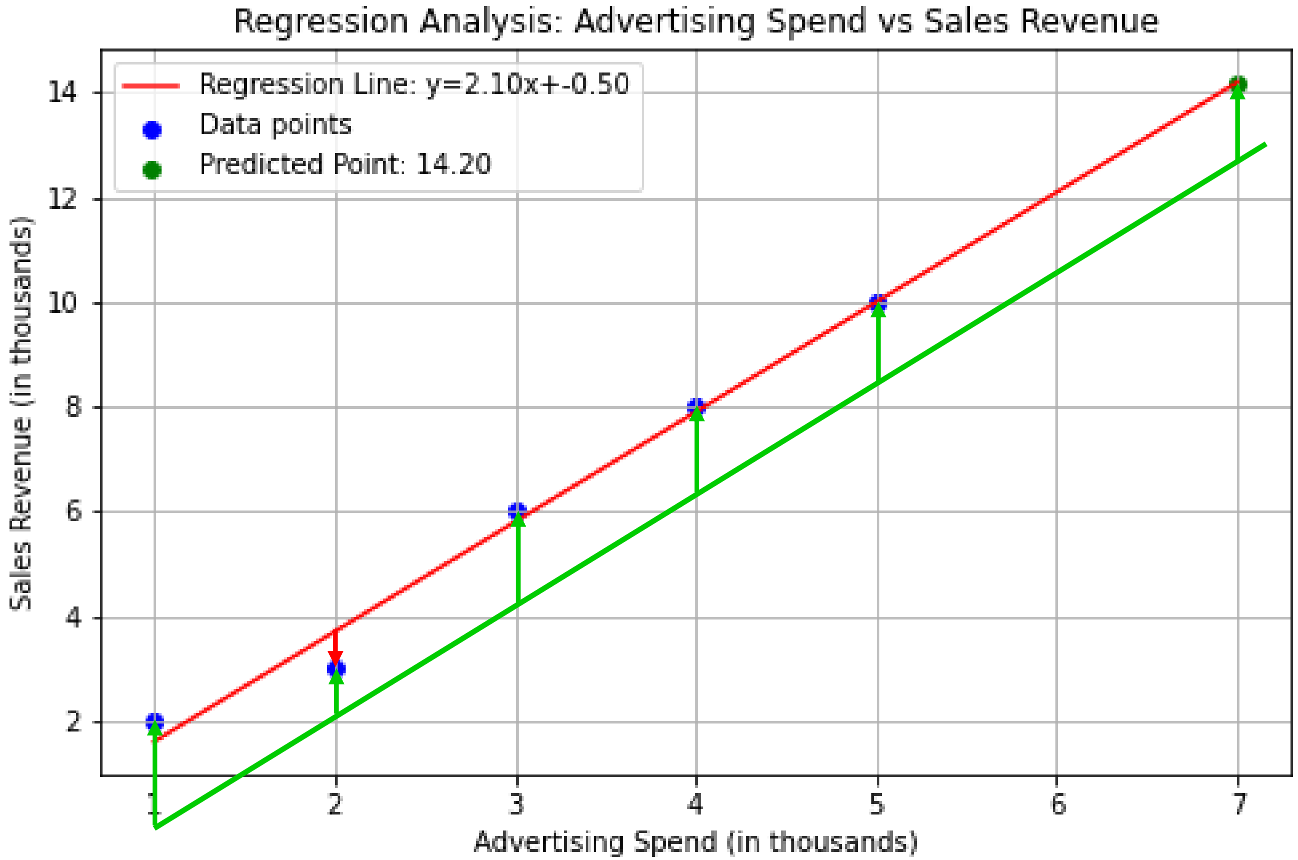


What about the error?

Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling, and finding the causal effect relationship between the variables.

Advertising Spend (in thousands)	Sales Revenue (in thousands)
1	2
2	3
3	6
4	8
5	10
7	14.20



What about the error?

Regression - Evaluate Model

Evaluating a regression model involves assessing its performance to understand how well it is predicting the outcomes. This typically includes calculating various statistical metrics that compare the predicted values produced by the model against the actual values in the test dataset. Common evaluation metrics for regression include:

- Mean Squared Error (MSE): The average of the squared differences between the predicted and actual values. MSE gives a rough idea of the magnitude of error (gives higher weight to larger errors). A lower MSE ($\mathrm{MSE} \geq 0$) value indicates a better fit between prediction and actual value.
- R-squared (R^2): Provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance. A higher R-squared ($0 \leq R^2 \leq 1$) value indicates a better fit between prediction and actual value.

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

Where:

- n is the number of observations.
- y_i is the actual value of the ith observation.
- $\hat{y}_i$ is the predicted value for the ith observation.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where:

- SS_{res} is the sum of squares of the residual errors ($\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$).
- $SS_{tot}$is the total sum of squares (total variation in the data) ($\sum_{i=1}^{n}(y_i - \bar{y})^2$).
- $\bar{y}$ is the mean value of y.

Polynomial Regression: Going Beyond Linear Relationships

Polynomial Regression is a form of regression analysis in which the relationship between the independent variable X and the dependent variable y is modeled as an nth degree polynomial. While Linear Regression is limited to linear relationships, Polynomial Regression can fit data with more complex trends by adding powers of the input variable as new variables.

Linear vs. Polynomial Regression

Linear Regression:

- It assumes a straight-line relationship between the variables.
- It has only two parameters, the slope and the intercept.
- It's best when the data trend is linear.

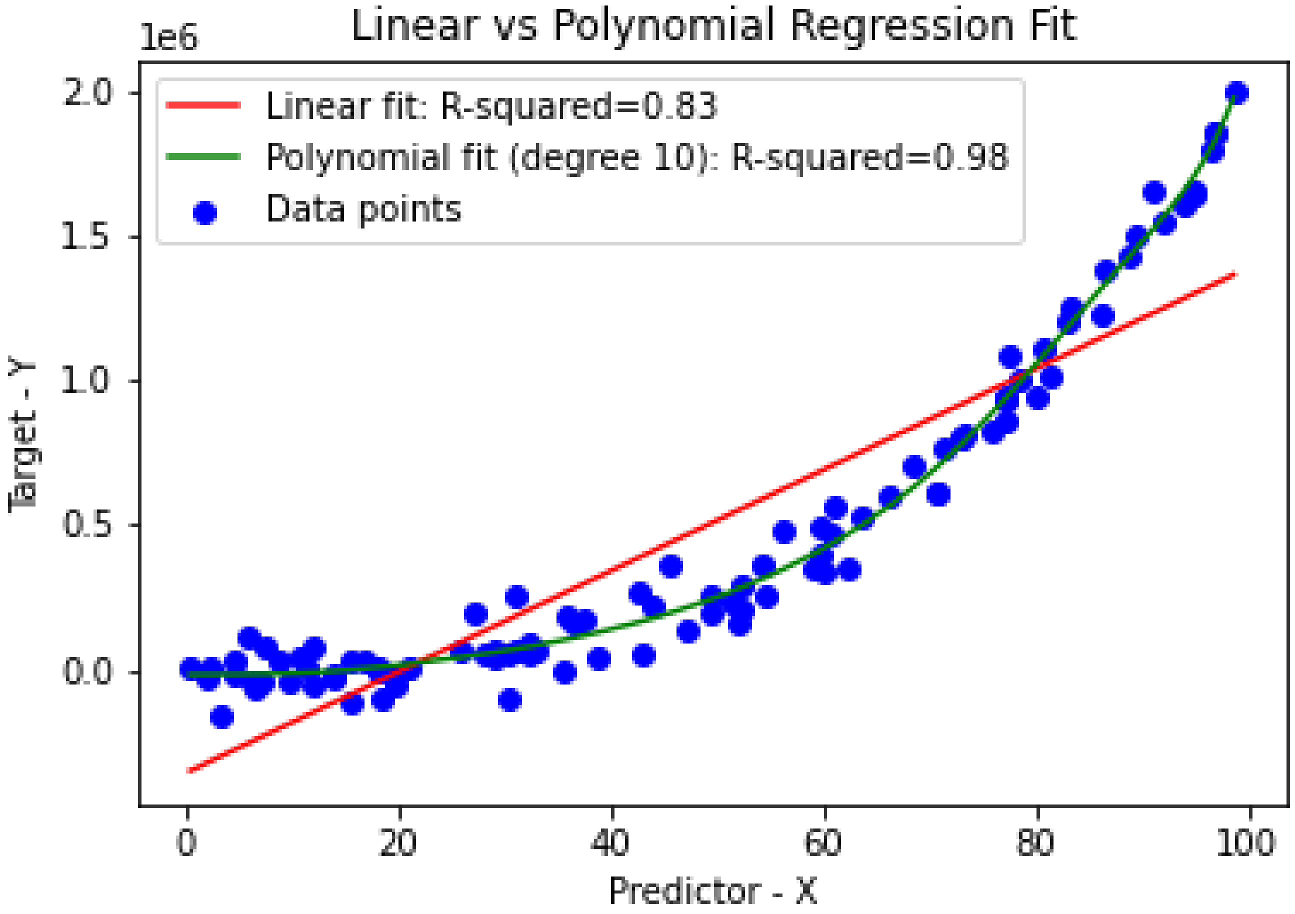
Polynomial Regression:

- It can capture the curvature in a dataset by including X^2, X^3, etc.
- It has more parameters than linear regression, one for each degree of the polynomial.
- It’s useful when data points show a curvilinear trend.

When to Use Polynomial Regression:

- Curved Trends: Use it when the data show a pattern or trend that isn't a straight line, indicating a potential polynomial relationship.
- Modeling Waves: It's ideal for phenomena that have peaks and troughs, as in seasonal effects or oscillatory systems.
- Fit Flexibility: When you need a model that can be more flexible in terms of fit, to capture the nuances in the data.

Polynomial Regression: Going Beyond Linear Relationships



The equation for the best polynomial fit of degree 10 is:

$$y = 1.60e - 02x + 8.94e - 03x^2 + 1.23e - 01x^3 + 9.76e - 01x^4 - 7.66e - 02x^5 + 2.75e - 03x^6 - 5.44e - 05x^7 + 6.14e - 07x^8 - 3.69e - 09x^9 + 9.16e - 12x^{10}$$