Machine learning 101

Agenda

- Timeshift
- Matplotlib
- Seaborn
- Machine Learning overview

Machine Learning 101

Overview

What is machine learning?

- Any model that uses stats to find pattern in data.

How can you use machine learning models?

ML Application

Financial Assets	Real variables	Operations
Treasury, stock price, commodities price, FX Extremely difficult Most direct to P&L	Revenue, macroeconomics indicator Medium level Need some interpretations to direct P&L	Sales, reports, risk, compliance Often overlooked See direct business impact

Types of ML

Supervised Learning

- Have a desired outcome
- Can be objectively measured
- Tree models, Neural Network

Unsupervised Learning

- Let the algo learn
- Can NOT be objectively measured

Reinforcement Learning

- System with rewards and penalties
- Agent, Environment, State, Action

Supervised Learning

- Supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known.
  - For example, a segment of text could have a category label, such as:
    - Spam vs. Legitimate Email
    - Positive vs. Negative Movie Review

Supervised Learning

- The network receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors.

- It then modifies the model accordingly.

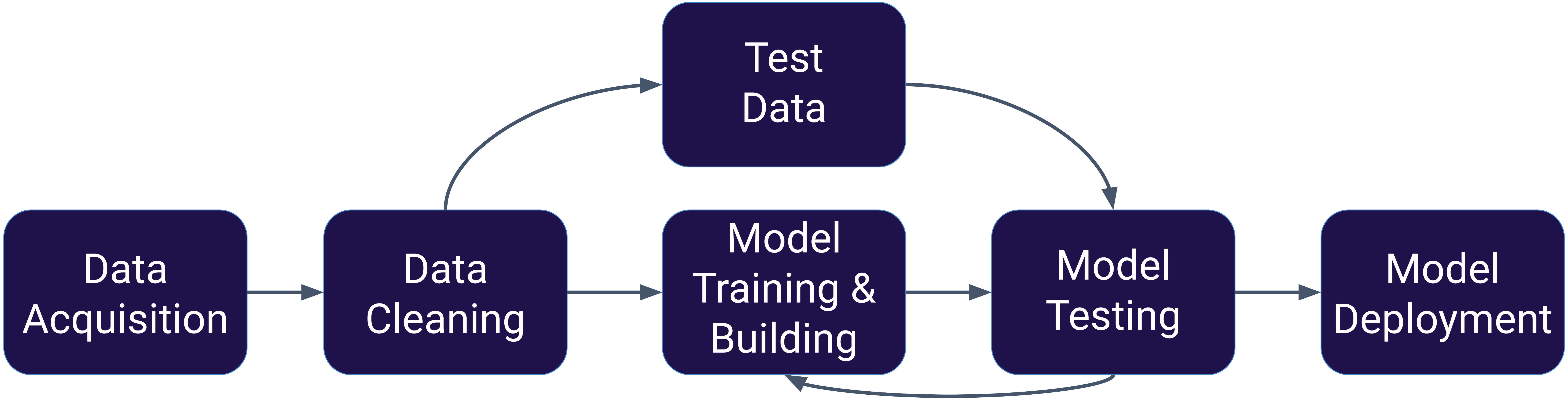
Source: Type source image link here

Supervised Learning

- Supervised learning is commonly used in applications where historical data predicts likely future events.

Source: Type source image link here

Supervised Learning



Supervised Learning

- What we just showed is a simplified approach to supervised learning, it contains an issue!
- Is it fair to use our single split of the data to evaluate our models performance?
- After all, we were given the chance to update the model parameters again and again.

Supervised Learning

- To fix this issue, data is often split into 3 sets
  - Training Data
    - Used to train model parameters
  - Validation Data
    - Used to determine what model hyperparameters to adjust
  - Test Data
    - Used to get some final performance metric

Machine Learning 101

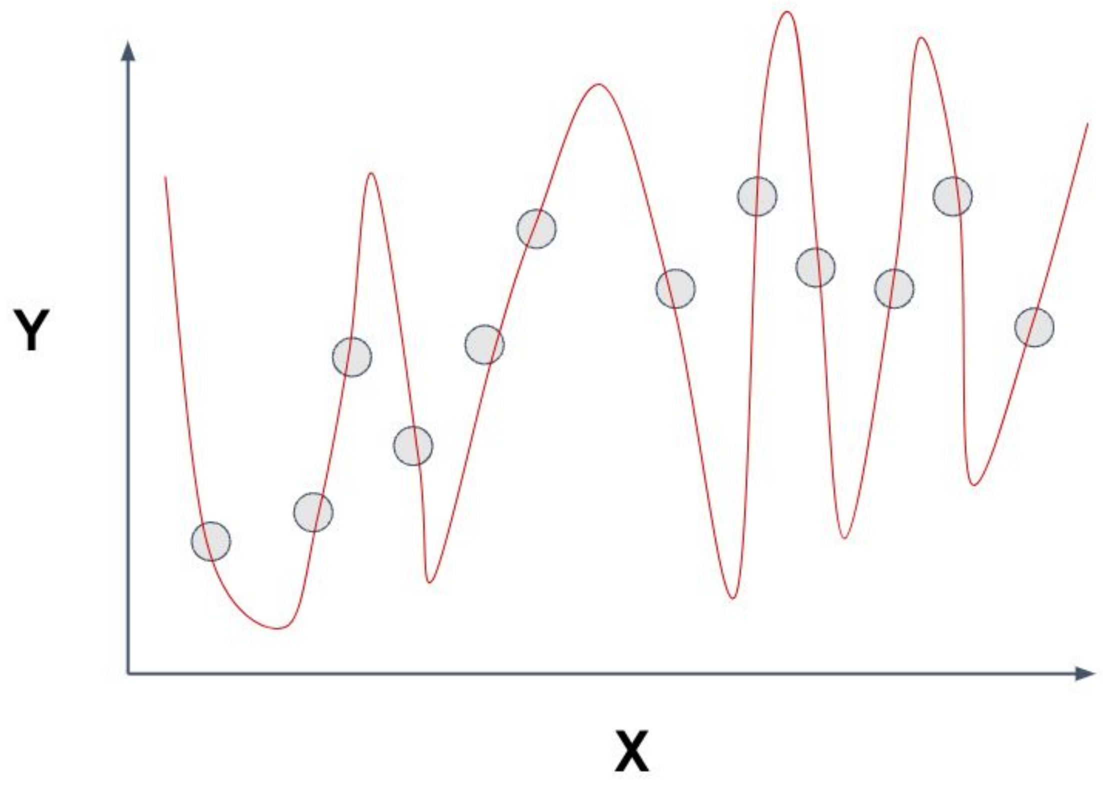
Overfitting and Underfitting

Fitting

- Now that we understand the full process for supervised learning, let's touch upon the important topics of overfitting and underfitting.

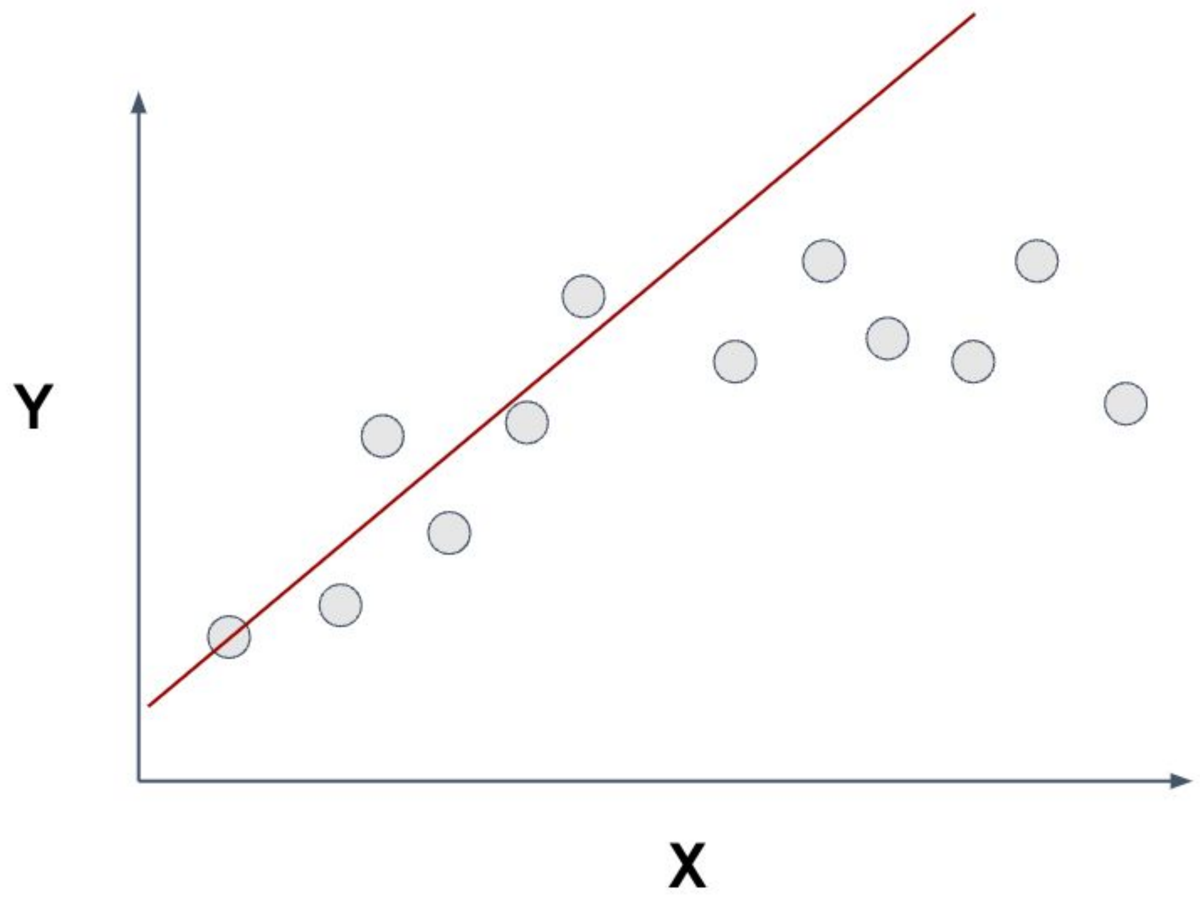
Fitting

- Overfitting
  - The model fits too much to the noise from the data.
  - This often results in **low error on training sets but high error on test/validation sets.**



Fitting

- **Underfitting**
  - Model does not capture the underlying trend of the data and does not fit the data well enough.
  - Low variance but high bias.
  - Underfitting is often a result of an excessively simple model.



Machine Learning 101

Classification Evaluation Performance

Model Evaluation

- The key classification metrics we need to understand are:
  - Accuracy
  - Recall
  - Precision
  - F1-Score

Model Evaluation

- Accuracy
    - Accuracy in classification problems is the number of correct predictions made by the model divided by the total number of predictions.

Model Evaluation

- Accuracy
    - For example, if the X_test set was 100 images and our model correctly predicted 80 images, then we have 80/100.
    - 0.8 or 80% accuracy.

- Accuracy is useful when target classes are well balanced

Model Evaluation

- Accuracy
  - Accuracy is **not a good choice with unbalanced classes**!
  - Imagine we had 99 images of dogs and 1 image of a cat.
- If our model was simply a line that always predicted dog we would get 99% accuracy!

Model Evaluation

- Recall
  - Ability of a model to find all the relevant cases within a dataset.
  - The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives.

Model Evaluation

- Precision
  - Ability of a classification model to identify only the relevant data points.
  - Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

Model Evaluation

- Recall and Precision
- Often you have a **trade-off** between Recall and Precision.
- While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant.

Model Evaluation

- F1-Score
- In cases where we want to find an optimal blend of precision and recall we can combine the two metrics using what is called the F1 score.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Model Evaluation

- F1-Score
  - We use the harmonic mean instead of a simple average because it punishes extreme values.
  - A classifier with a precision of 1.0 and a recall of 0.0 has a simple average of 0.5 but an F1 score of 0.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Confusion Matrix

		predicted condition	
	total population	prediction positive	prediction negative
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**

Machine Learning 101

Regression Evaluation Performance

Evaluating Regression

- Let's discuss some of the most common evaluation metrics for regression:
- Mean Absolute Error
- Mean Squared Error
- Root Mean Square Error

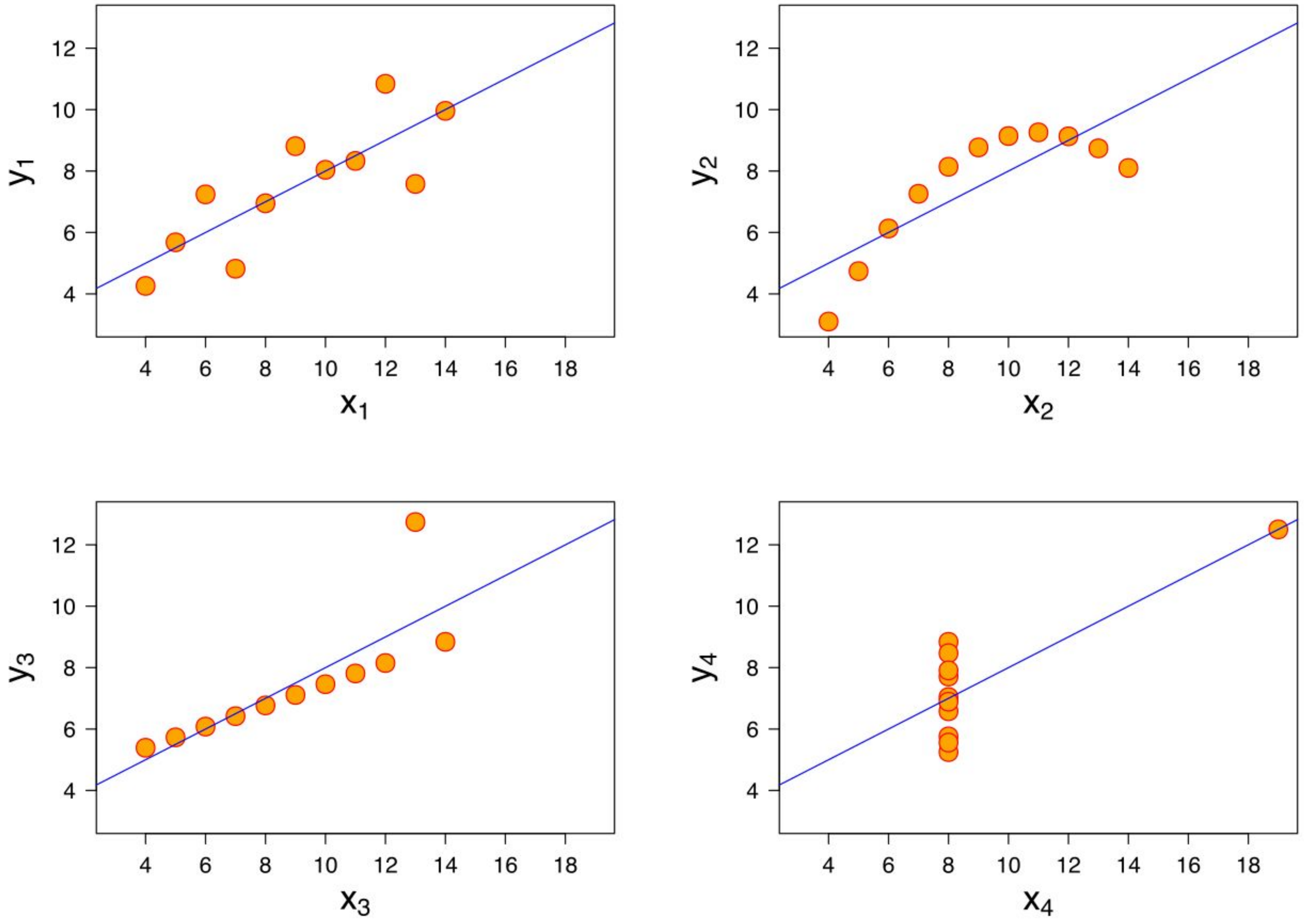
Evaluating Regression

- Mean Absolute Error (MAE)
    - This is the mean of the absolute value of errors.
    - Easy to understand

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

Evaluating Regression

- MAE won't punish large errors however.



Evaluating Regression

- Root Mean Square Error (RMSE)
- This is the root of the mean of the squared errors.
- Most popular (has same units as y)

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

Machine Learning 101

Unsupervised Learning

Unsupervised Learning

- There are certain tasks that fall under unsupervised learning:
- Clustering
- Anomaly Detection
- Dimensionality Reduction

Unsupervised Learning

- Clustering
  - Grouping together unlabeled data points into categories/clusters
  - Data points are assigned to a cluster based on similarity

Unsupervised Learning

- Anomaly Detection
  - Attempts to detect outliers in a dataset
  - For example, fraudulent transactions on a credit card.