

A financial fraud predictive AI model integrated with Spring Boot.

Predictive Model for Bank Industry

Table Of Contents

What is Machine Learning

Key Concepts

Supervised Learning Process

Introduction to Anomaly
Detection

Anomaly Detection Use Cases

Example

Dataset Description

Data Preparation

How Logistic Regression Works

Logistic Regression Implementation

How Support Vector Machine Works

SVM Implementation

Fraud in Banking Industry

In 2024, more than half banks reported an increase in business fraud



Increase of frauds

In 2024, check fraud losses in the Americas reached nearly \$21 billion, and 70% of U.S. financial institutions reported check fraud.



Check Fraud

Over half report increasing investment in third-party fraud prevention.



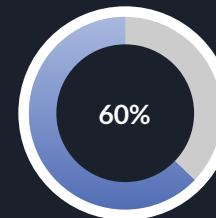
Increase of Investment

- Thompson Reuters, Deloitte, and the Federal Trade Commission (FTC)
<https://www.alloy.com/state-of-fraud-benchmark-report-2024#component-marketo-embed>
- SEC. (n.d.). <https://www.sec.gov/files/fy24-oiad-sar-objectives-report.pdf>



Preview of Results

Predictive model classifies 60% of frauds over 24 million bank transactions



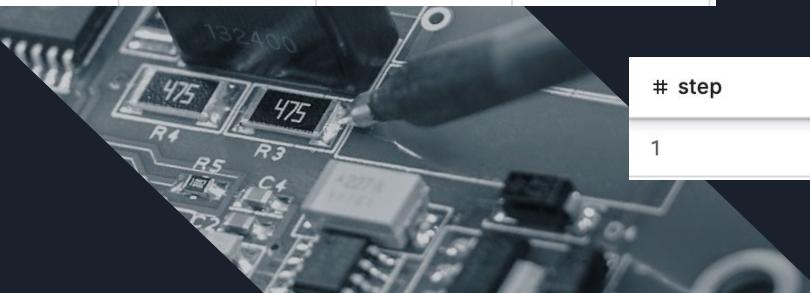
Detection of Frauds

What is the best approach?

# step	# type	# amount	# nameDest	# isFraud
1	PAYMENT	9839.64	M1979787155	0
1	PAYMENT	1864.28	M2044282225	0
1	TRANSFER	181.0	C553264065	1
1	CASH_OUT	181.0	C38997010	1
1	PAYMENT	11668.14	M1230701703	0
1	PAYMENT	7817.71	M573487274	0
1	PAYMENT	7107.77	M408069119	0
1	PAYMENT	7861.64	M633326333	0
1	PAYMENT	4024.36	M1176932104	0
1	DEBIT	5337.77	C195600860	0



# step	# type	# amount	# nameDest
1	PAYMENT	5885.56	M1804441305

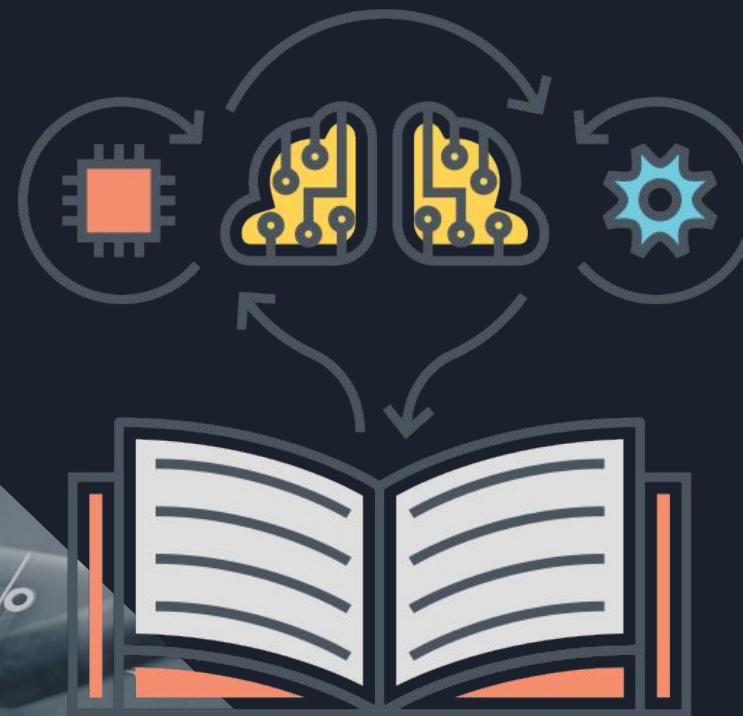




What is the best approach?



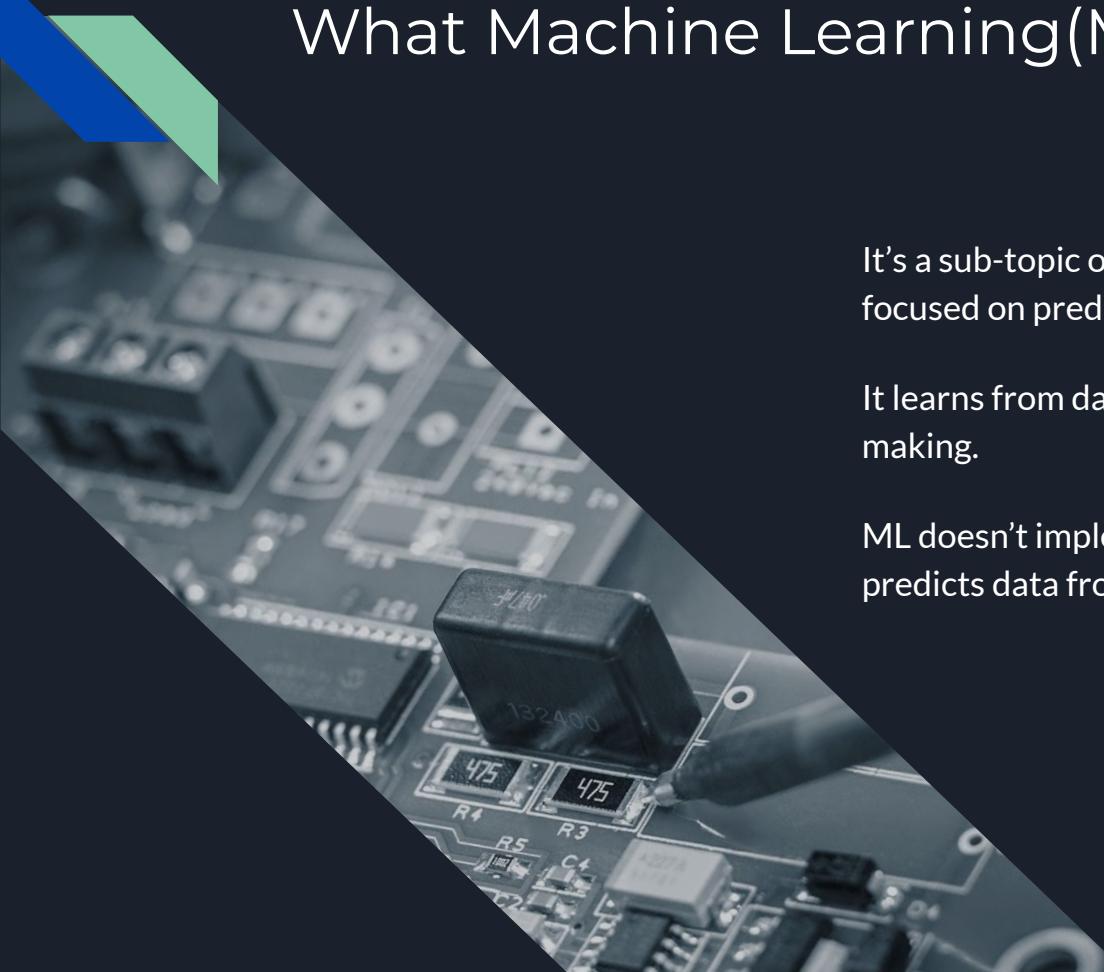
What is the best approach?



Machine learning



What Machine Learning(ML) is



It's a sub-topic of Artificial Intelligence field, which is focused on predicting data, through information patterns.

It learns from data to enhance automation of decision making.

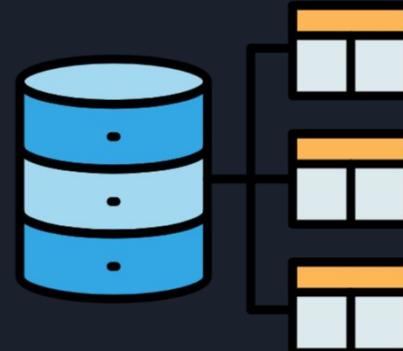
ML doesn't implement hard-coded rules. Instead, it predicts data from unseen information.



Data

01

The more data, the better.

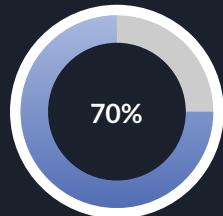


02

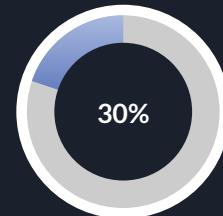
Data format can be numbers, text, images, sounds.

03

It is divided into training (used to teach the model) and test (used to evaluate the model).



Train



Test

Model

01

Mathematical representation that learns from data.

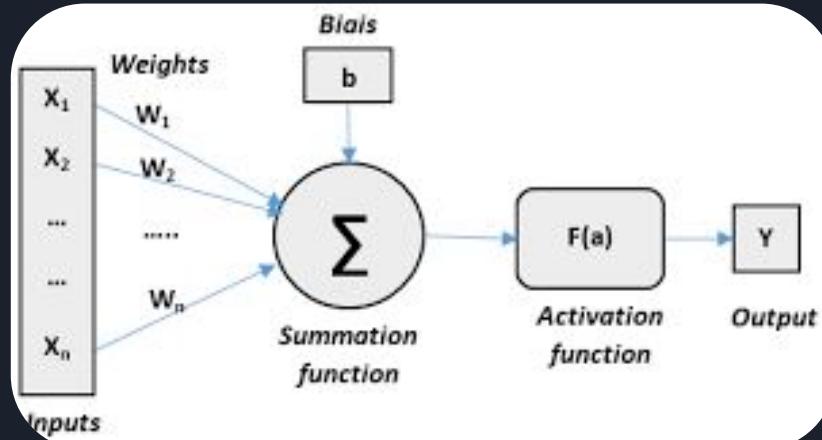
02

Models capture data patterns and relationships to make predictions and decisions.

03

Typical models are:

- Logistic Regression
- Support Vector Machine
- Neural Networks, etc.





Training & Prediction

01

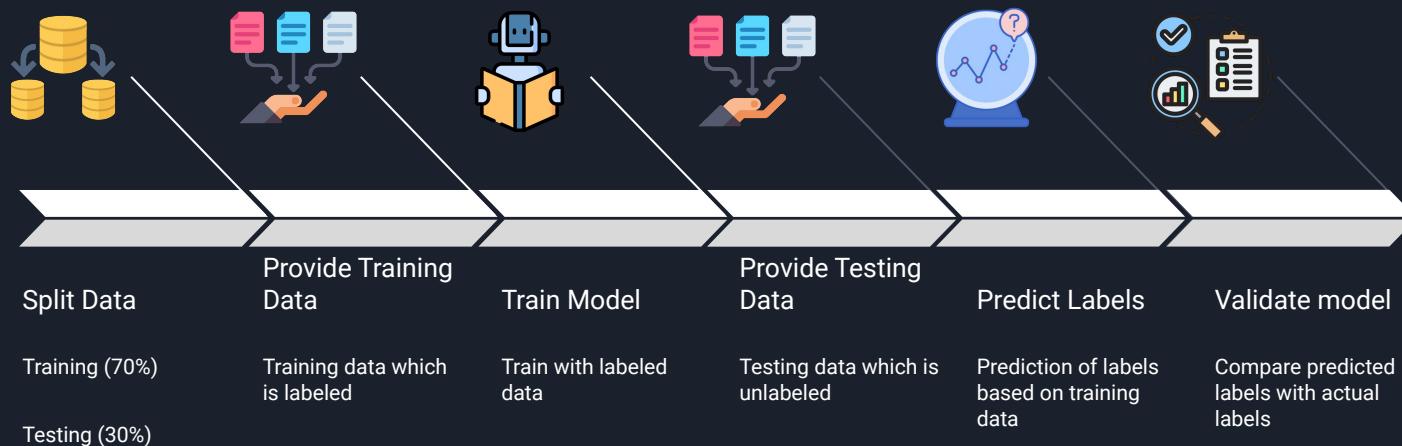
Training is teaching mathematical model to recognize patterns in data.

02

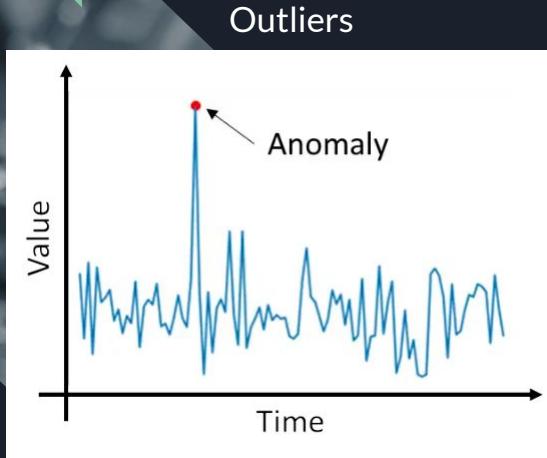
Data can be labeled(supervised learning) or unlabeled(unsupervised learning).



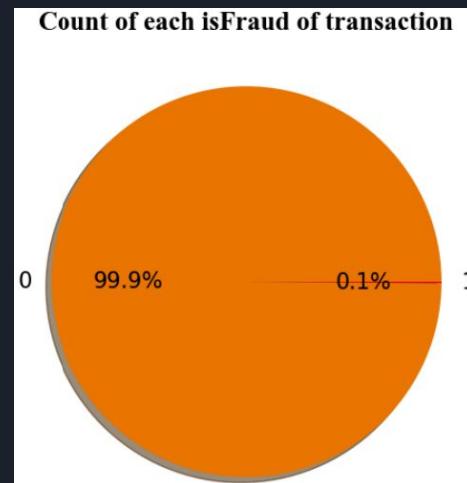
Supervised Learning Process



Anomaly Detection



Machine learning model used to detect outliers in extreme unbalanced datasets, usually big enough to consider high-demand computational processing.



Unbalanced dataset over 24 million

Some Anomaly Detection Use Cases



01

Financial Fraud:

Identifying unusual spending patterns on credit cards to detect fraudulent transactions.



02

Healthcare Monitoring:

Identifying unusual vital signs in patient data that could indicate a medical emergency.



03

Retail Sales Analysis:

Identifying sudden spikes or drops in sales for specific products that could indicate issues with pricing or demand.

Dataset Description

“PaySim is a financial simulator that simulates mobile money transactions based on an original dataset.”

The dataset has 24 million financial records.

E. A. Lopez-Rojas, A. Elmir, and S. Axelsson. "PaySim: A financial mobile money simulator for fraud detection". In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus. 2016

Dataset Fields

- **step:** Represents a unit of time in the real world, with 1 step equating to 1 hour.
- **type:** Transaction types include CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER.
- **amount:** The transaction amount in the local currency.
- **nameOrig:** The customer initiating the transaction.
- **oldbalanceOrg:** The initial balance before the transaction.
- **newbalanceOrig:** The new balance after the transaction.
- **nameDest:** transaction's recipient customer.
- **oldbalanceDest:** The initial recipient's balance before the transaction.
- **newbalanceDest:** The new recipient's balance after the transaction.
- **isFraud:** Identifies transactions conducted by fraudulent agents aiming to deplete customer accounts through transfers and cash-outs.

Example of Anomaly Detection

Data Preparation

Feature Selection:

Shapiro-Wilk test to determine null hypothesis - > not normal distribution. > 0.05 confidence level:

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$(a_1, \dots, a_n) = \frac{m^\top V^{-1}}{\|V^{-1}m\|} = (m^\top V^{-1}V^{-1}m)^{1/2}$$

V : Variance Covariance Matrix

m : x

Data Preparation

Feature Selection:

Pearson Correlation: Correlation between two normal distributed variables.

Coefficient correlation: -1 to 1

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

σ : Standard Deviation

COV : Variance Covariance Matrix

Data Preparation

Feature Selection:

Spearman Correlation: Correlation between two non-normal distributed variables.

Coefficient correlation: -1 to 1

$$r_s = \rho [R[X], R[Y]] = \frac{\text{cov} [R[X], R[Y]]}{\sigma_{R[X]} \sigma_{R[Y]}},$$

$\text{cov} [R[X], R[Y]]$: Covariance of ranked variables

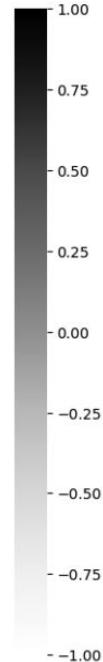
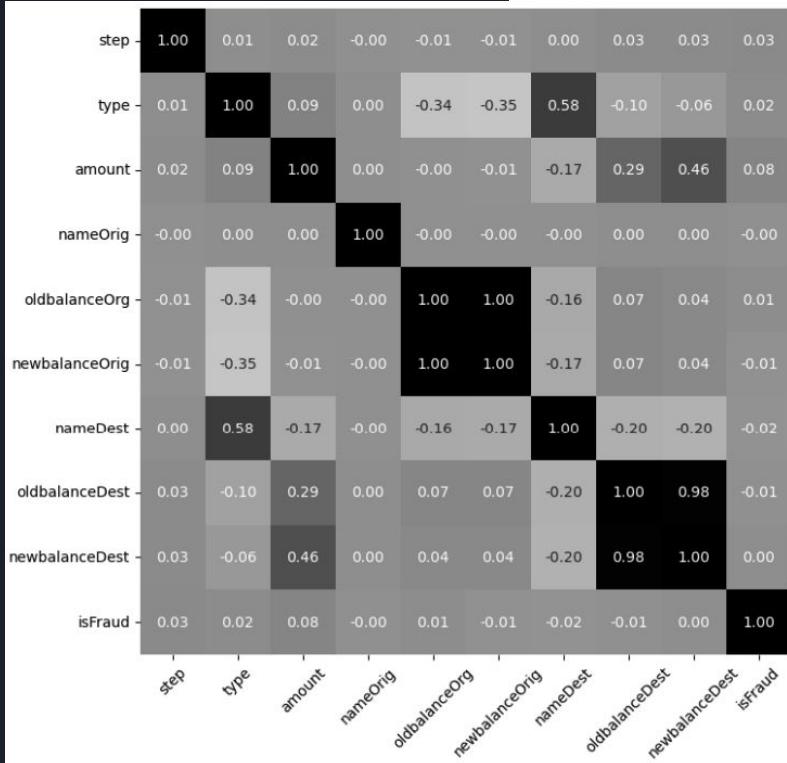
$\sigma_{R[X]} \sigma_{R[Y]}$: Standard Deviation



Example of Anomaly Detection

Data Preparation

Pearson
Correlation
Matrix



Spearman
Correlation
Matrix

Data Preparation

Test Hypothesis

Logistic Regression: Relationship between two variables. Only applicable when dependent variable (y) is categorical, and independent variable (x) is continuous.

p-Value: 0 - 0.05 (confidence threshold)

$$p(x) = \frac{1}{1 + e^{-(\beta_0 - \beta_1 x)}}$$

β : Weight optimization parameter

Data Preparation

Test Hypothesis

Chi-Square: Relationship between two variables. Only applicable when dependent variable (y) is categorical, and independent variable (x) is also categorical.

p-Value: 0 - 0.05 (confidence threshold)

$$\chi^2 = \sum_{i=1}^n \frac{O_i^2}{E_i} - N.$$

O_i : Number of observations of variable i

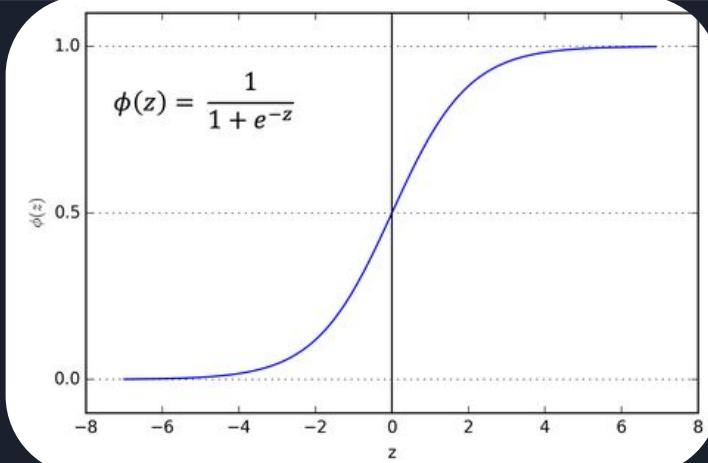
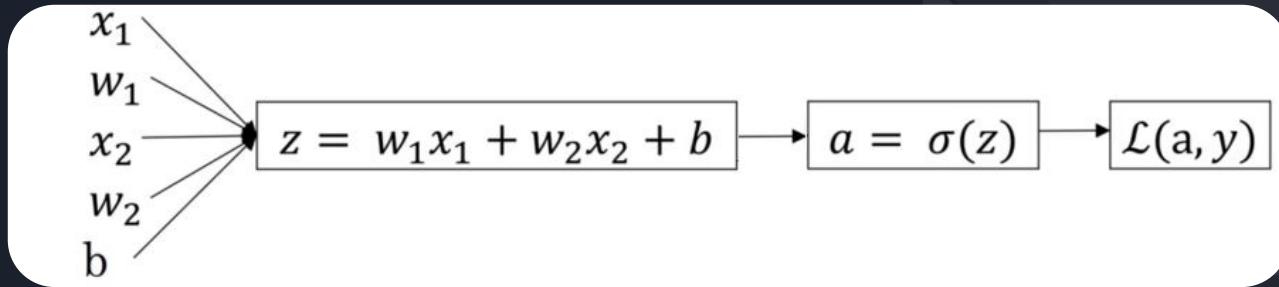
N : Total number of observations

E_i : Expected number of type i

Example of Anomaly Detection

How Logistic Regression Works

Classification Function



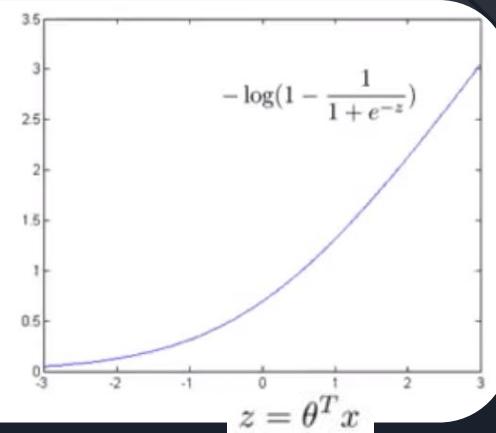
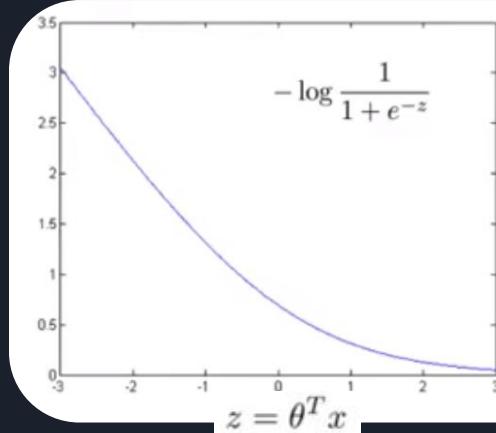
Sigmoid Function

Example of Anomaly Detection

How Logistic Regression Works

Cost Optimization Function

$$\text{Cost}(h_\theta(x), y) = -y \log \frac{1}{1 + e^{-\theta^T x}} - (1 - y) \log(1 - \frac{1}{1 + e^{-\theta^T x}})$$



Example of Anomaly Detection

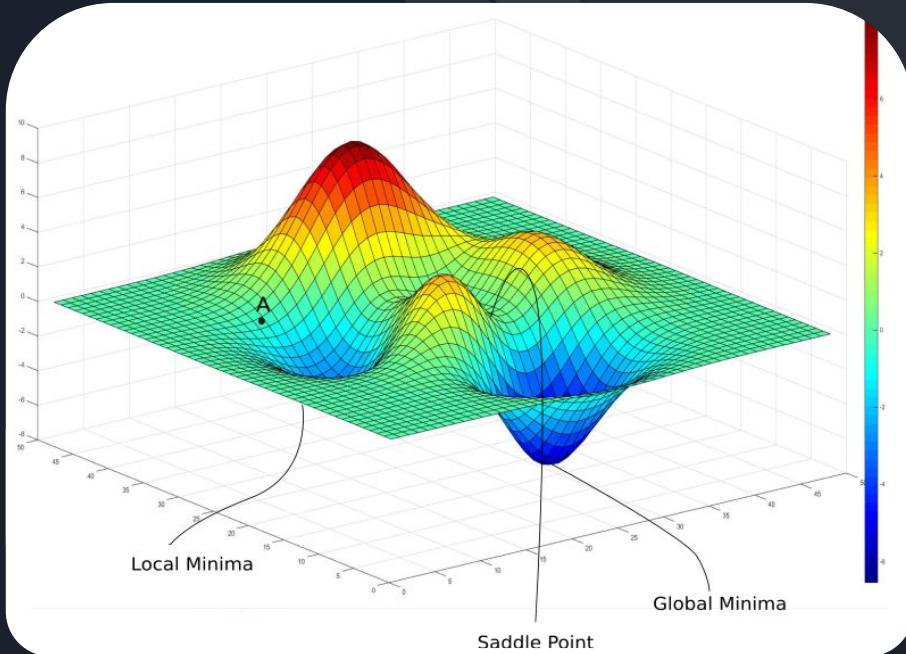
How Logistic Regression Works

Cost Optimization Function

Repeat until convergence {

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

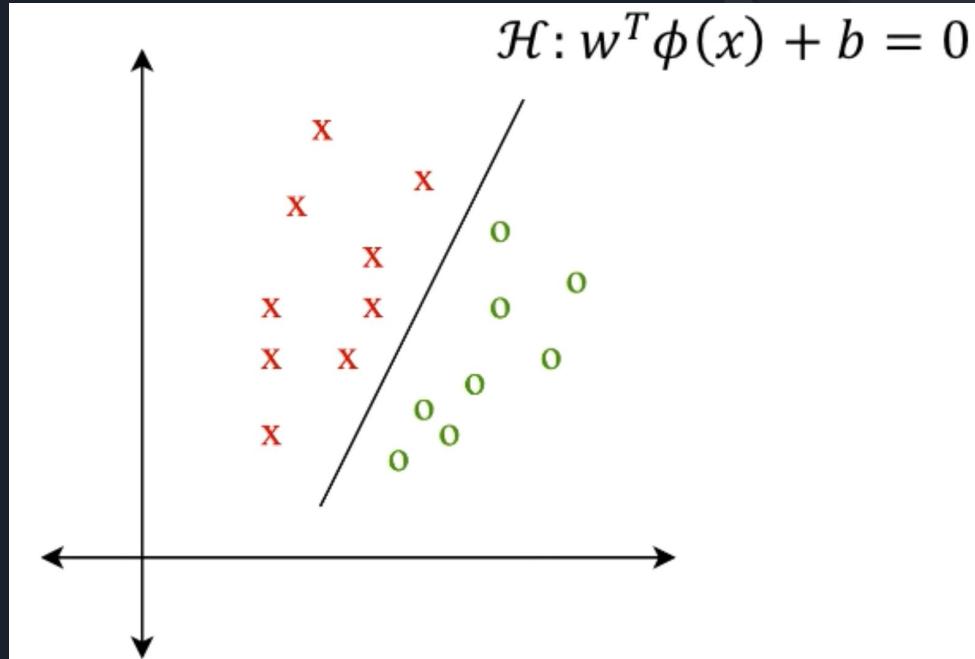
}



Example of Anomaly Detection

How Support Vector Machine Works

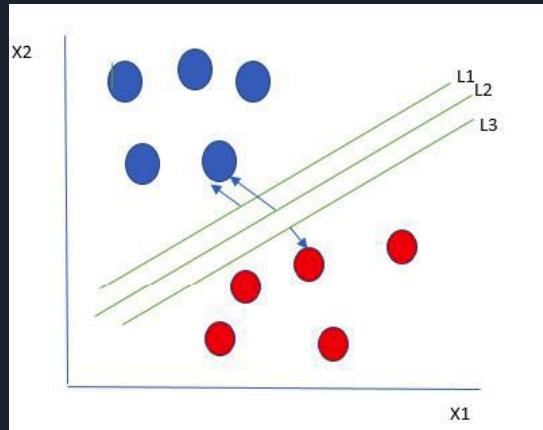
Classification Function: Hyperplane classifier



Example of Anomaly Detection

How Support Vector Machine Works

Optimization Function



$$w^* = \arg \max_w \left[\min_n d_{\mathcal{H}}(\phi(x_n)) \right]$$

Optimization function: Max distance of the hyperplane of the nearest point from the hyperplane

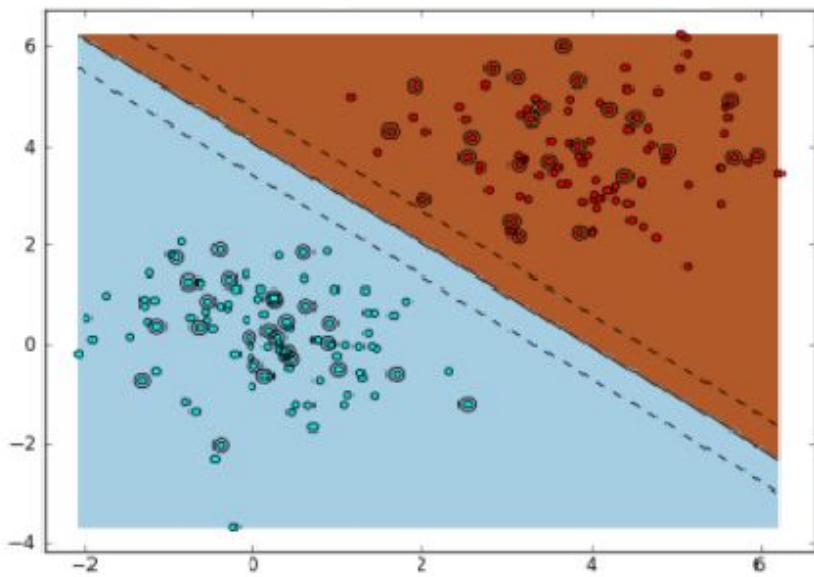
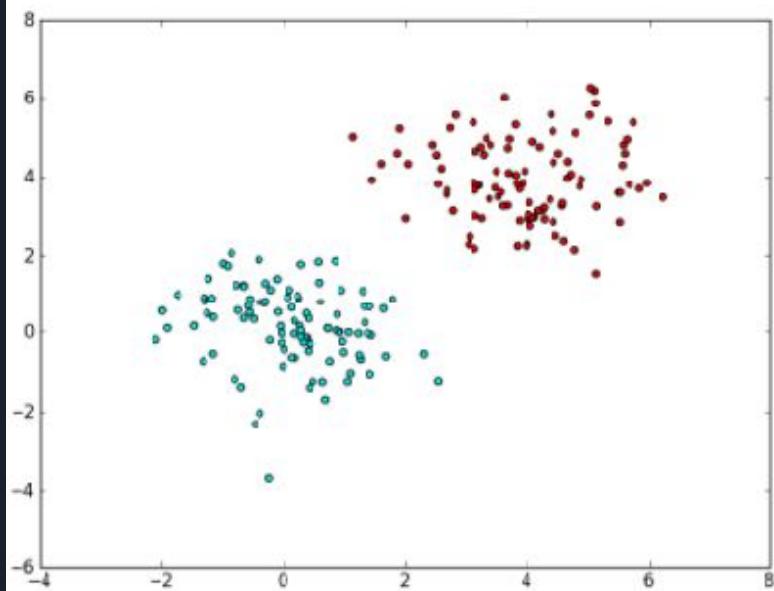
$$d_{\mathcal{H}}(\phi(x_0)) = \frac{|w^T \phi(x_0) + b|}{\|w\|_2}$$

Function distance of points from hyperplane

Example of Anomaly Detection

How Support Vector Machine Works

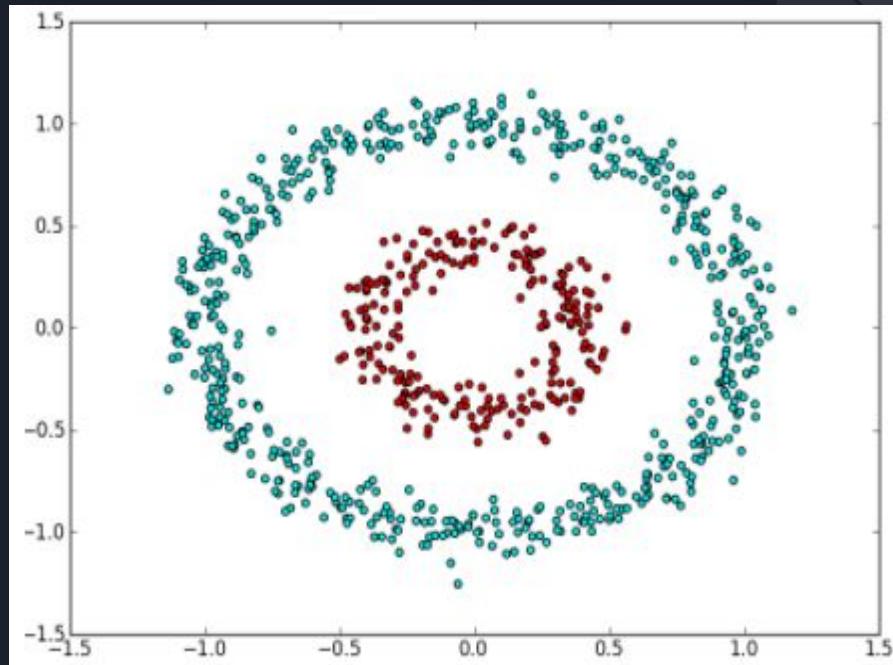
Kernel



Example of Anomaly Detection

How Support Vector Machine Works

Kernel



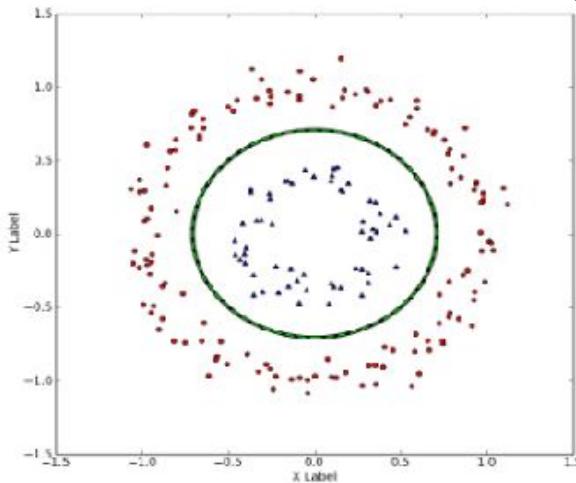
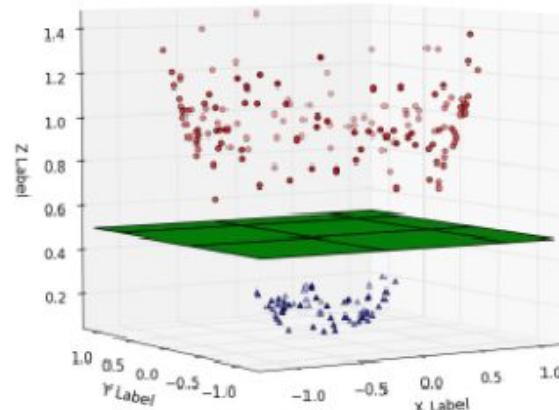
Example of Anomaly Detection

How Support Vector Machine Works

Kernel

Lagrange Multipliers

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3 .$$



Hilbert Space



Example

Implementation

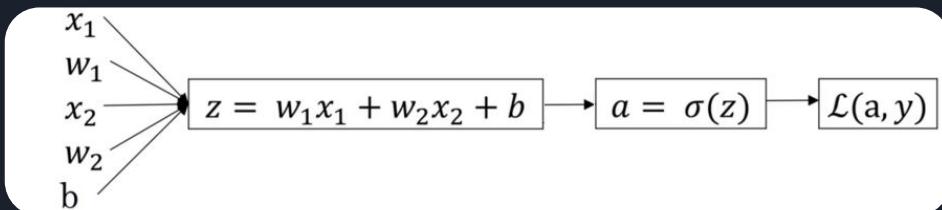
- 01 Inferential Analysis.
- 02 Data Preparation.
- 03 Implementation of Logistic Regression
- 04 Implementation of Support Vector Machine
- 05 Spring Boot Integration



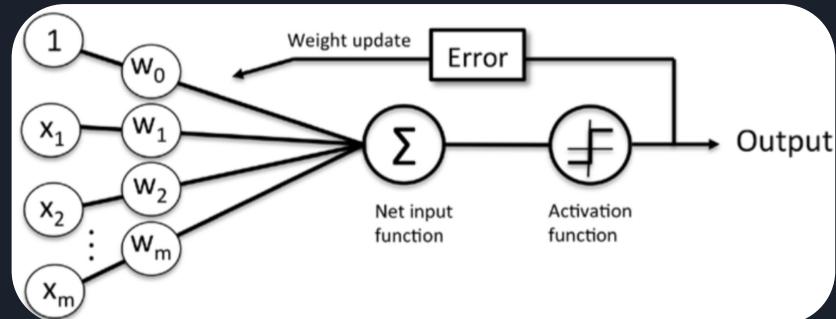
Logistic regression v/s Neural Network

Bonus

Logistic Regression Function



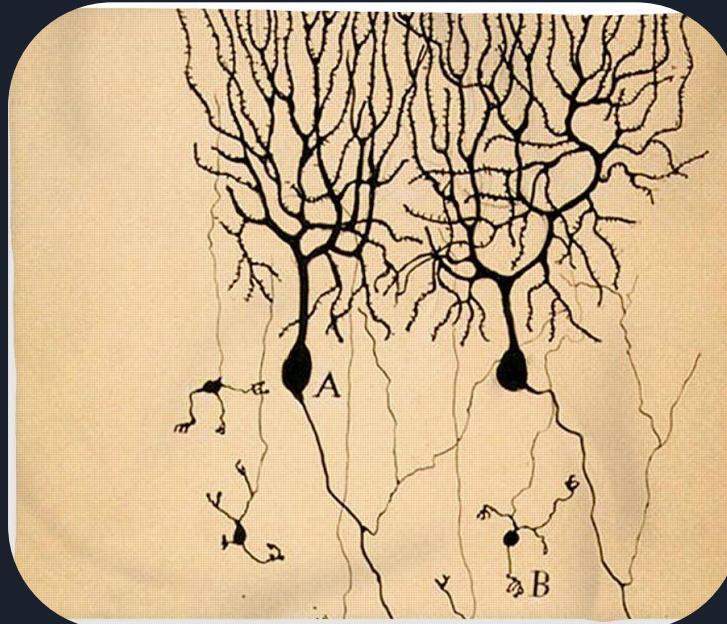
Neuron Model





Logistic regression v/s Neural Network

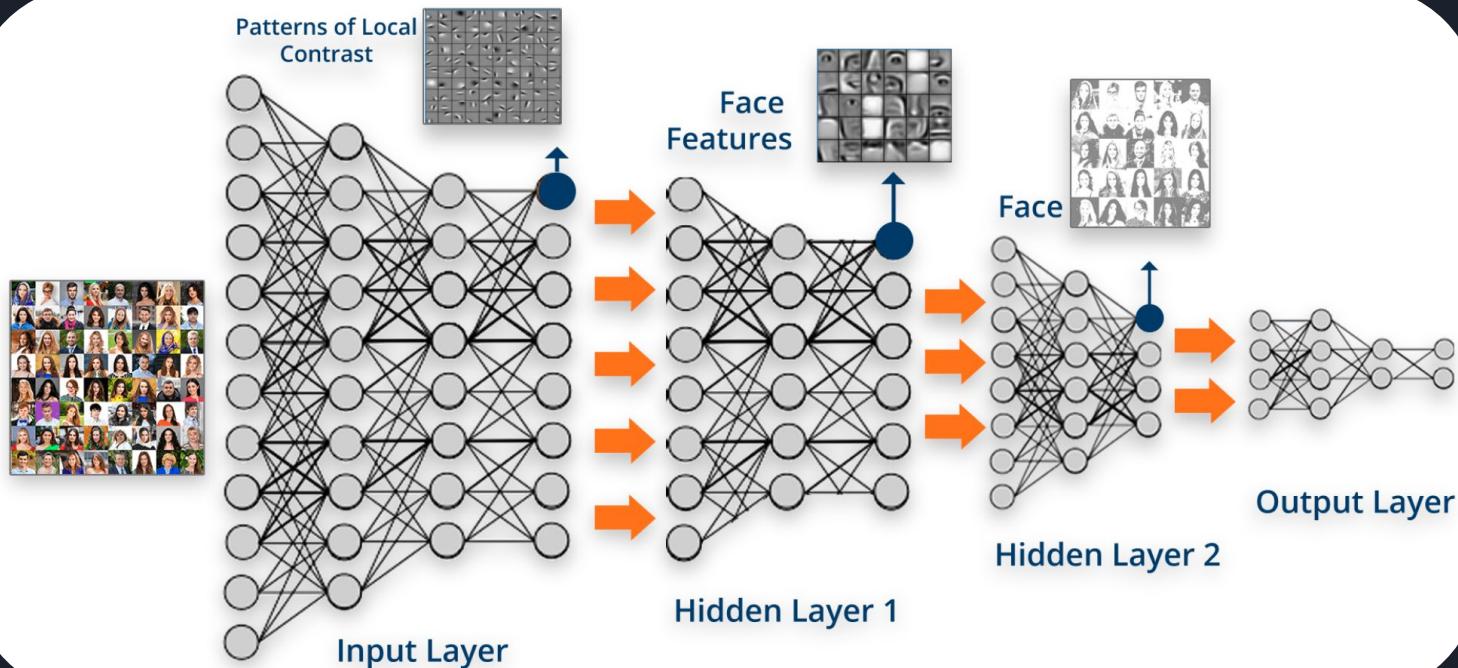
Bonus



Modelo Neuronal - Santiago Ramon y Cajal (1906)

Logistic regression v/s Neural Network

Bonus





Bibliography

- Lopez-Rojas, E., Elmir, A., & Axelsson, S. (2016). *PaySim: A financial mobile money simulator for fraud detection*. In *28th European Modeling and Simulation Symposium, EMSS, Larnaca* (pp. 249-255). Dime University of Genoa.
- Mahesh, B. (2020). *Machine learning algorithms-a review*. *International Journal of Science and Research (IJSR).[Internet]*, 9(1), 381-386.
- Hilal, W., Gadsden, S. A., & Yawney, J. (2022). *Financial fraud: a review of anomaly detection techniques and recent advances*. *Expert systems With applications*, 193, 116429.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., et al. (2020). *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*. *Nature Methods*, 17, 261-272. doi:10.1038/s41592-019-0686-2
- Kramer, O., & Kramer, O. (2016). *Scikit-learn. Machine learning for evolution strategies*, 45-53.
- Komer, B., Bergstra, J., & Eliasmith, C. (2019). *Hyperopt-sklearn. Automated Machine Learning: Methods, Systems, Challenges*, 97-111.
- Sarkar, T. (2022). *GPU-Based Data Science for High Productivity*. In *Productive and Efficient Data Science with Python: With Modularizing, Memory profiles, and Parallel/GPU Processing* (pp. 299-326). Berkeley, CA: Apress
- SEC. (n.d.). <https://www.sec.gov/files/fy24-oiad-sar-objectives-report.pdf>

Thank you!

