CLOUD COMPUTING

DATS -6450_12

PHISHING WEBSITE DETECTION USING AWS CLOUD SERVICES

Group 5

Aakash Singh Sivaram

Hema Chandra Puchakayala

Pramod Krishnachari

Professor: Melo, Walcelio

1st December, 2025

SCOPE

GOAL:

Design, build, and deploy a cloud-based machine-learning system on AWS that automatically classifies websites as phishing or legitimate based on URL, domain, and security-related attributes.

Impact:

Provides an automated, cloud-native system to identify malicious websites, reducing manual analysis and enhancing threat detection accuracy.

Dataset

UCI Phishing Websites Dataset

Description:

- 11k+ records with preprocessed data and features.
- 30 input attributes derived from URL structure, domain features, and SSL certificate data.
- Already cleaned and normalized - no manual feature engineering required.

Data Flow

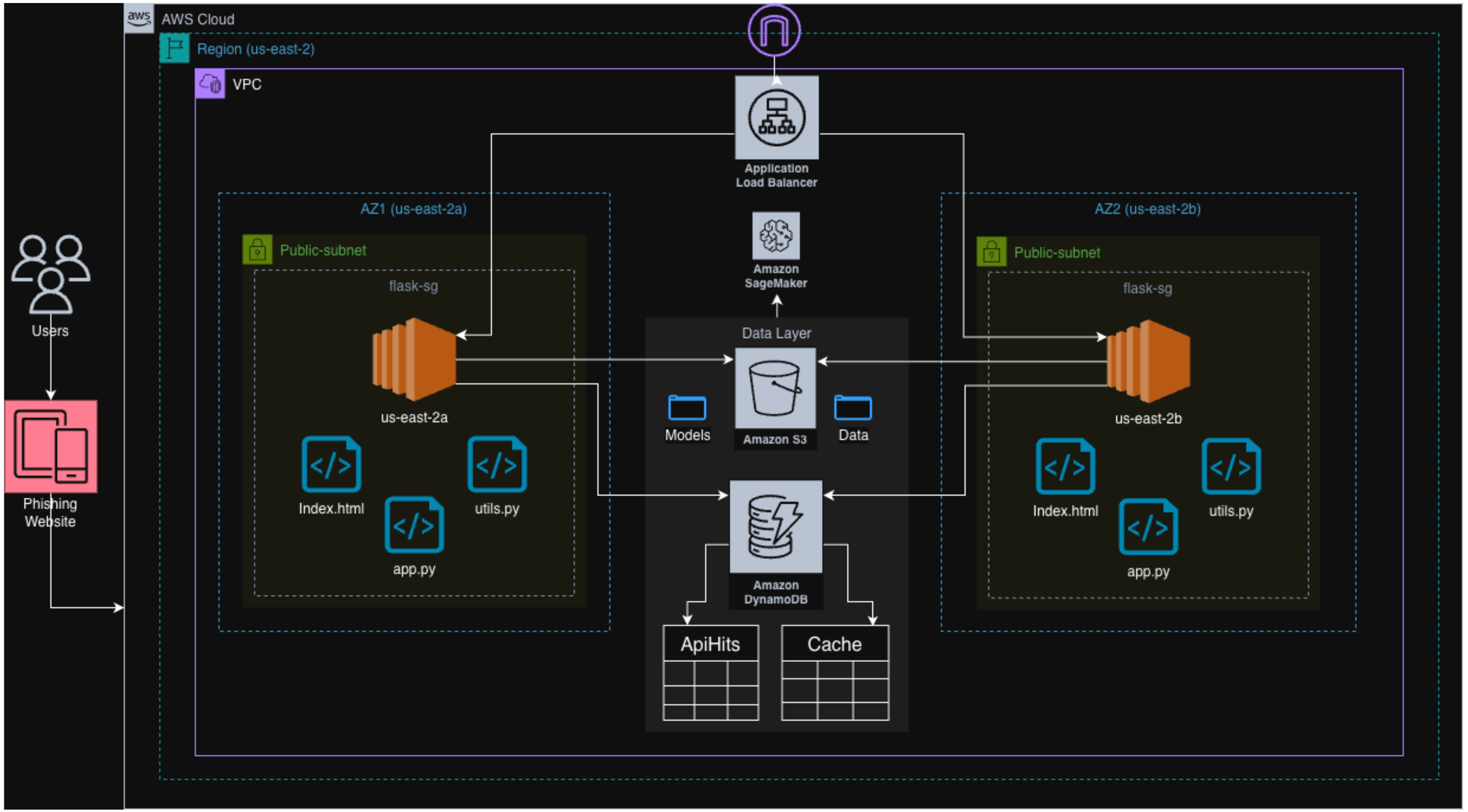
Model Training & Data Preparation Flow

1. Phishing Website Dataset → Amazon S3
2. S3 → Amazon SageMaker (training environment)
3. SageMaker → S3 (model artifacts)

Real time inference

1. Users visit a phishing-website URL
2. Users → Application Load Balancer
3. ALB → EC2 instances (multi-AZ)
4. EC2 → DynamoDB (Cache lookup)
5. EC2 → S3 (if additional resources needed)
6. EC2 → Model Inference
7. EC2 → DynamoDB (write ApiHits + update Cache)
8. EC2 → ALB → Users (response)

AWS Architecture

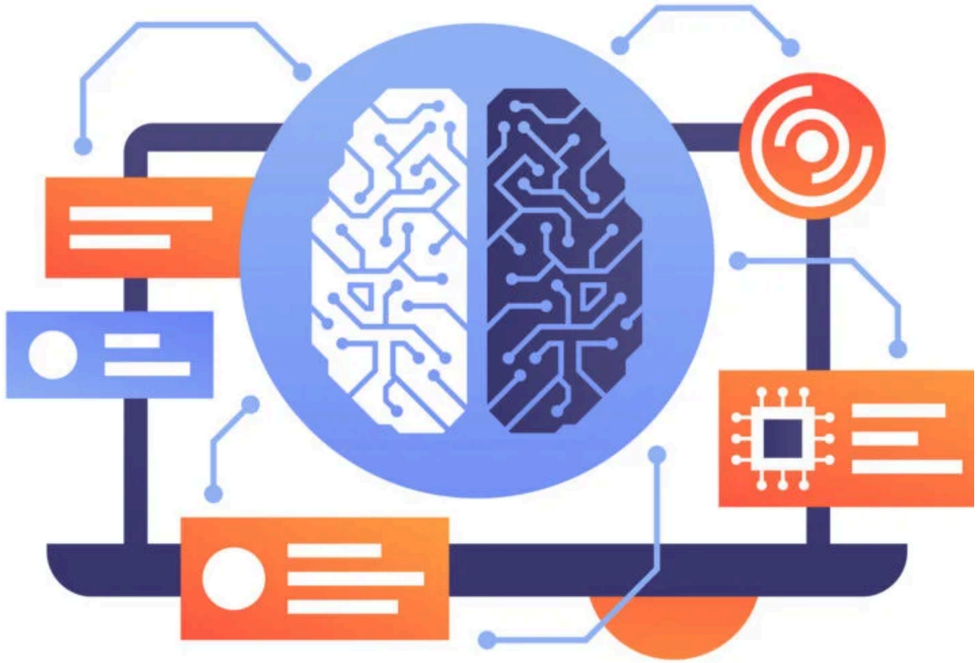


AWS Feature

Purpose	AWS Service	Description
Data Storage	Amazon S3	Stores raw dataset and trained model artifacts.
Metrics Exploration & EDA	Jupyter labs in SageMaker	Analyze and visualize Metrics
Model Training & Tuning	Amazon SageMaker	Train ML models
Model Evaluation	SageMaker Evaluation Jobs	Evaluation Metrics

Modelling

Model Type	Why Consider
SVM	Excellent for binary classification with clear margins
Multilayer Perceptron	Captures complex nonlinear patterns
XGBoost	Handles heterogeneous features and captures complex nonlinear relationships.
Random Forest	Robust to overfitting and works well with high-dimensional, noisy data
Decision Tree	Provides clear decision rules and handles mixed feature types.
Auto Encoder	Useful for feature extraction, denoising, and anomaly detection.



Results and Conclusion

	ML Model	Train Accuracy	Test Accuracy
2	Multilayer Perceptrons	0.866	0.867
3	XGBoost	0.867	0.864
1	Random Forest	0.816	0.817
0	Decision Tree	0.813	0.813
5	SVM	0.802	0.801
4	AutoEncoder	0.380	0.359

XGBoost Performs the best among other models in classifying phishing websites