Nguyễn Chí Thanh-64TTNT2

2251262639

**The Robustness of Machine Learning Models Using MLSecOps : A Case Study On Delivery ServiceForecasting**

Bối cảnh của nghiên cứu là sự gia tăng của thương mại điện tử đã thúc đẩy nhu cầu về dịch vụ giao hàng và tạo ra lượng dữ liệu lớn cho việc dự báo và tối ưu hóa hoạt động[**The rise of e-commerce and shifting consumer behavior have led to a surge in demand for delivery services . This trend has generated a wealth of data that holds valuable information about patterns and trends in service demand. Utilizing this data effectively has become crucial for companies to forecast and anticipate future delivery requirements. Known as Delivery Service Forecasting , this process plays a fundamental role in optimizing and managing modern delivery service operations.**]. Dự báo dịch vụ giao hàng (Delivery Service Forecasting) sử dụng các mô hình và kỹ thuật dựa trên dữ liệu để dự đoán nhu cầu tương lai, thời gian giao hàng, và phân bổ nguồn lực nhằm tối ưu hóa hậu cần, cải thiện sự hài lòng của khách hàng và hợp lý hóa hoạt động. Thuật toán Random Forest được coi là một công cụ mạnh mẽ cho mục đích dự báo này, nhờ khả năng xử lý các tập dữ liệu lớn, nhận diện các mẫu phức tạp và đưa ra dự báo chính xác[**Delivery service forecasting utilizes data-driven models and techniques to anticipate future demand, predict delivery times, and allocate resources efficiently. Its aim is to optimize logistics, improve customer satisfaction, and streamline operations. By analyzing historical data, market trends, and external factors, precise forecasts can be generated to guide decision making and enhance the overall performance of delivery services**

**The random Forest algorithm has emerged as a powerful tool for forecasting.Its capacity to process extensive datasets, discern complicated patterns, and produce accurate forecasts makes it a perfect choice for improving delivery operations and boosting customer satisfaction]**

Tuy nhiên, một thách thức đáng kể là tính dễ bị tổn thương của các mô hình ML trước các cuộc tấn công đối kháng (adversarial attacks)[**The integration of Machine Learning Operations (MLOps) and security, known as Machine Learning Security Operations (MLSecOps), has emerged as a crucial aspect in the contemporary field of machine learning. MLSecOps combines security practices with Machine Learning Operations (MLOps) to enhance the security, robustness, and reliability of machine learning models, mitigating vulnerabilities and ensuring trustworthy deployment in production environments[14].**

**Recent studies have demonstrated the superiority of Random Forest in accurately predicting service outcomes, such as estimated delivery times or delays. The algorithm's capability to handle large and complex datasets, capture nonlinear relationships, and mitigate the impact of outliers and noise has been widely acknowledged. Although Random Forest is a good algorithm for delivery service forecasting, there is a risk of encountering boundary attacks and backdoor attacks. Although Random Forest is a good algorithm for delivery service forecasting, there is a risk of encountering boundary attacks and backdoor attacks. One method to enhance the robustness of machine learning models is by utilizing the Adversarial Robustness Toolbox**]

Nghiên cứu này tập trung vào hai loại tấn công:

(Boundary attack): thay đổi nhẹ các điểm dữ liệu để đánh lừa mô hình trong khi vẫn ở gần dữ liệu gốc[**The robustness of machine learning models against adversarial attacks, such as boundary and backdoor attacks, is a significant concern. Boundary attacks subtly modify data points to mislead models while remaining near the original data . Backdoor attacks, on the other hand, involve a malicious actor subtly manipulating the training data to control the model's output . Both expose the vulnerabilities of machine learning models, warranting defenses such as adversarial training.**

**The Adversarial Robustness Toolbox (ART) has emerged as a comprehensive library to foster robust machine learning. ART provides interfaces for various adversarial attacks and defenses. It supports the implementation and combination of diverse defense methods, enabling a tailored approach to enhance model robustness]-[1]**

(Backdoor attack): kẻ tấn công thao túng dữ liệu huấn luyện một cách tinh vi để kiểm soát đầu ra của mô hình[1].

Để tăng cường tính bảo mật, mạnh mẽ và khả năng phục hồi của mô hình ML, nghiên cứu đã áp dụng MLSecOps (Machine Learning Security Operations)1.... MLSecOps tích hợp các hoạt động bảo mật vào vòng đời của mô hình ML, nhằm xây dựng các hệ thống kiên cường trước các mối đe dọa.... Công cụ MLSecOps được sử dụng trong nghiên cứu này là Adversarial Robustness Toolbox (ART).... ART là một thư viện cung cấp các công cụ toàn diện để bảo vệ mô hình ML khỏi tấn công đối kháng, bao gồm các kỹ thuật phát hiện, tạo và đánh giá các ví dụ đối kháng, cũng như các phương pháp phòng thủ như huấn luyện đối kháng (adversarial training)[1].... ART được IBM duy trì hoạt động[**A machine learning platform is a comprehensive solution for efficiently training, deploying, and managing machine learning models. It simplifies the model building process, automates tasks, and guarantees scalability, thereby aiding organizations in effectively utilizing data-driven insights.** ].

Nghiên cứu đã thực hiện một thí nghiệm áp dụng MLSecOps và ART vào việc dự báo dịch vụ giao hàng sử dụng thuật toán Random Forest[**We conducted an experiment on the application of MLSecOps in delivery service forecasting using the Random Forest algorithm presented through a comprehensive case study. We will assess how the amalgamation of MLSecOps principles and the utilization of the Adversarial Robustness Toolbox can fortify model dependability and precision, even amidst harmful attacks such as boundary and backdoor incursions. Our study spans the model's formation,appraises its robustness against prospective attacks, and refines the model in adherence to MLSecOps principles.**]Dữ liệu sử dụng là dữ liệu giao hàng gốc giai đoạn 2021-2023 từ Kaggle, chứa 9 biến và 991,230 dòng[**The dataset used in this work includes the original delivery data for the period 2021-2023 containing nine variables, as described in Table 1. Data were gathered from Kaggle. The dataset consisted of 991,230 rows.**]. Quy trình MLSecOps bao gồm các bước như thu thập và tiền xử lý dữ liệu, phát triển và xác thực mô hình có chú trọng bảo mật, triển khai an toàn, giám sát liên tục và cải tiến lặp lại[**The machine learning security operation (MLSecOps) process includes tasks such as data collection and preprocessing to ensure data quality and security, model development and validation with a focus on security aspects and robust model machine learning, secure and robust deployment of machine learning models, continuous monitoring and maintenance to detect and mitigate potential threats, and iterative improvements for security and performance]**.

Thí nghiệm được tiến hành theo kịch bản ba giai đoạn để đánh giá tính mạnh mẽ”

Đánh giá hiệu suất mô hình trong điều kiện bình thường[**The boundary attack experiment involves testing the accuracy, precision, f1-score, and recall, which are also displayed on the ROC graph under normal conditions. The next stage involves testing accuracy, precision, f1-score, and recall, which are also displayed on the ROC graph during a boundary attack condition. The final stage involves testing accuracy, precision, f1-score, and recall, also shown on the ROC graph, when the machine learning model has been trained to be more robust using the Adversarial Robustness Toolbox tool.**]-[2]

Đánh giá hiệu suất mô hình trong điều kiện bị tấn công [2]

Đánh giá hiệu suất mô hình sau khi được huấn luyện để mạnh mẽ hơn sử dụng công cụ ART[2]

Kết quả thí nghiệm cho thấy tính dễ bị tổn thương của mô hình trước các cuộc tấn công[**Condtion Performance model**

**Accuracy Precision Recall F1-Score**

**Normal condition 93,5 % 90,7 % 97 % 93,7 %**

**Backdoor attack condition 50,5 % 50,5 % 100 % 67 %**

**robustness model machine learning**

**93,5 % 90,7 % 97 % 93,7 %**]-[3]

Trong điều kiện tấn công đường biên (Boundary attack): Độ chính xác của mô hình Random Forest đã giảm đáng kể từ 93.5% xuống còn 50.5% F1-Score và độ chuẩn xác (precision) cũng bị ảnh hưởng đáng kể[3].... Tuy nhiên, sau khi kết hợp ART trong giai đoạn huấn luyện để tăng tính mạnh mẽ, hiệu suất của mô hình đã được khôi phục về trạng thái ban đầu, đạt độ chính xác 93% và giá trị AUC ROC 0.9217....

Trong điều kiện tấn công cửa hậu (Backdoor attack): Mô hình cũng gặp suy giảm hiệu suất[**Condtion Performance model**

**Accuracy Precision Recall F1-Score**

**Normal condition 93,5 % 90,7 % 97 % 93,7 %**

**Backdoor attack condition 50,5 % 50,5 % 100 % 67 %**

**robustness model machine learning**

**93,5 % 90,7 % 97 % 93,7 %**

**V. CONCLUSION**

**The results of this experiment highlight the vulnerability of machine learning models to boundary and backdoor attacks. However, the implementation of techniques such as the adversarial robustness toolbox proved to be effective in bolstering the models' strength and protecting them against potential breaches.**]-[4]Sau khi huấn luyện với ART, mô hình đã phục hồi khả năng hoạt động như trong điều kiện bình thường, đạt độ chính xác 93.5% và giá trị AUC ROC 0.9320....

Kết luận của nghiên cứu là kết quả thí nghiệm đã làm nổi bật tính dễ bị tổn thương của mô hình ML trước tấn công đường biên và cửa hậu[4]. Tuy nhiên, việc triển khai các kỹ thuật như sử dụng ART đã chứng minh hiệu quả trong việc tăng cường sức mạnh và bảo vệ mô hình[4]....

Điều này nhấn mạnh tầm quan trọng của việc huấn luyện các mô hình ML để tăng tính mạnh mẽ trước nhiều loại tấn công mạng [**Similarly, the model faced a backdoor assault, leading to a decline in its performance. However, post-training with the adversarial robustness toolbox, the model was able to restore its performance capability to the levels seen in standard conditions.. The importance of training machine learning models to increase their strength and robustness against a variety of cyberattacks has become clear. Utilizing tools such as the adversarial robustness toolbox can significantly boost their performance, even in situations of attacks. Despite this, we must remain vigilant and keep the model and its defenses updated, as attackers may develop new strategies to interfere with their operations.**]-[5]. Mặc dù vậy, cần duy trì cảnh giác và cập nhật mô hình cùng các biện pháp phòng thủ do khả năng phát triển các chiến lược tấn công mới của kẻ xấu[5].