УДК 004.491

Нейтрализация полезной нагрузки в моделях машинного обучения

Менисов А.Б., Ситников И.Д. (Военно-космическая академия имени А.Ф. Можайского)
Научный руководитель – кандидат технических наук, старший преподаватель Менисов А.Б.
Военно-космическая академия имени А.Ф. Можайского)

Введение. Распространение открытых хранилищ моделей машинного обучения создаёт риски внедрения вредоносного программного обеспечения в предварительно обученные модели. Через внедрение вредоносной полезной нагрузки в модели машинного обучения можно осуществлять несанкционированный доступ к системам, манипулировать выводами моделей и выполнять вредоносные действия на целевой инфраструктуре. Это значительно расширяет возможности атакующих и создает новые угрозы безопасности для систем, использующих искусственный интеллект.

Разработан метод скрытого внедрения вредоносной нагрузки в модели машинного обучения с использованием различных форматов сериализации, технологий и архитектур хранения весов, а также стеганографии. Созданы скрипты для внедрения и извлечения полезной нагрузки из моделей PyTorch, TensorFlow, ONNX, HDF5, Keras [1].

Основная часть. Процесс исследования был построен поэтапно:

1. изучение структуры хранения весов моделей различных архитектур машинного обучения;
2. разработка методов внедрения полезной нагрузки в тензоры;
3. создание вредоносных файлов с использованием разработанных методов;
4. проведение экспериментов по внедрению и извлечению полезной нагрузки из моделей PyTorch, TensorFlow, ONNX, HDF5, Keras;
5. анализ результатов экспериментов и оценка эффективности разработанных методов.

Экспериментальное исследование включало анализ структуры хранения весов моделей, разработку различных методов для внедрения полезной нагрузки в тензоры, создание вредоносных файлов, вариаций исполнения вредоносного кода для автоматического выполнения полезной нагрузки при загрузке модели [2].

Выводы. Демонстрируется возможность внедрения вредоносной нагрузки в предварительно обученные модели машинного обучения различных архитектур без значительного влияния на их работу [3]. Метод масштабируем и применим к различным технологиям MLOPS. Теоретически подтверждена возможность скрытного внедрения вредоносного ПО через модели машинного обучения. Практическое применение заключается в применении программного комплекса с целью тестирования на проникновение систем, использующих модели машинного обучения. Перспективы развития направления связаны с разработкой более эффективных методов защиты от подобных атак.

Список использованных источников:

1. Шила Т., Гнана Падмеш С.К., Локеш В. Предотвращение и обнаружение отравляющих атак в веб-приложениях машинного обучения //Международная конференция по коммуникации, сетям и вычислениям. – Чам: Springer Nature Switzerland, 2022. – С. 128-

2. Вуд А.М. Защитная стратегия обнаружения целевых состязательных атак с отравлением в моделях обнаружения вредоносных программ, обученных машинному

обучению. – 2021.
3. Лин Дж. и др. Модели атак машинного обучения: состязательные атаки и атаки по отравлению данных //препринт arXiv arXiv:2112.02797. – 2021