Исследователи обнажили новое оружие хакеров для взлома ML-систем.

Недавнее исследование от компании Trail of Bits – это компания, специализирующаяся на кибербезопасности и разработке инструментов для обеспечения безопасности программного обеспечения и блокчейн-технологий. Компания предоставляет услуги по аудиту кода, уязвимостям и кибербезопасности для различных организаций и проектов.

 Trail of Bits также разрабатывает инструменты и решения для обнаружения и предотвращения кибератак, а также помогает клиентам улучшать безопасность своих приложений и систем. Компания активно внедряет технологии и методологии, которые помогают защищать информацию и активы от киберугроз." data-html="true" data-original-title="Trail of Bits" >Trail of Bits выявило новую технику атаки на модели машинного обучения (ML) под названием «Sleepy Pickle». Эта атака использует популярный формат Pickle, который применяется для упаковки и распространения моделей машинного обучения, и представляет серьёзный риск для цепочки поставок, угрожая клиентам организаций.

Исследователь безопасности Боян Миланов отметил, что Sleepy Pickle представляет собой скрытую и новую технику атаки, направленную на саму модель машинного обучения, а не на основную систему.

Формат Pickle широко используется библиотеками ML, такими как PyTorch, и может быть использован для выполнения произвольного кода путём загрузки файла Pickle, что создаёт потенциальную угрозу.

В документации Hugging Face – это компания, занимающаяся разработкой искусственного интеллекта и специализирующаяся на области обработки естественного языка (Natural Language Processing, NLP). Она была основана в 2016 году и стала популярной благодаря своим инновационным библиотекам и инструментам для NLP.

 Одним из наиболее известных продуктов компании является Hugging Face Transformers – открытая библиотека для обучения, использования и разработки моделей глубокого обучения в области NLP. Эта библиотека предоставляет широкий спектр предобученных моделей, которые могут быть использованы для различных задач, таких как классификация текста, извлечение информации и машинный перевод." data-html="true" data-original-title="Hugging Face" >Hugging Face рекомендуется загружать модели только от проверенных пользователей и организаций, полагаться лишь на подписанные коммиты, а также использовать форматы TensorFlow или Jax с механизмом автоконверсии «from_tf=True» для дополнительной безопасности.

Атака Sleepy Pickle работает путём вставки вредоносного кода в файл Pickle с помощью таких инструментов, как Fickling, и доставки этого файла на целевую систему через различные методы, включая атаки типа Adversary-in-the-Middle (AitM) – это метод хакерской атаки, при котором злоумышленник встраивается между двумя устройствами или системами в сети, чтобы перехватывать, модифицировать или вмешиваться в передачу данных между ними.

 Этот метод атаки позволяет злоумышленнику получать доступ к конфиденциальной информации, внедрять вредоносное программное обеспечение или манипулировать данными между участниками обмена информацией. Атаки типа AitM представляют серьезную угрозу для безопасности данных и часто используются в целях кражи личных данных, выявления уязвимостей и других злонамеренных целей." data-html="true" data-original-title="AitM" >AitM, фишинг, компрометацию цепочки поставок или использование уязвимостей системы.

При десериализации на системе жертвы вредоносный код выполняется и изменяет модель, добавляя в неё бэкдоры, контролируя выходные данные или подделывая обрабатываемую информацию до её возвращения пользователю. Таким образом, злоумышленники могут изменять поведение модели, а также манипулировать данными входа и выхода, что может привести к вредоносным последствиям.

Гипотетическая атака может привести к генерации вредоносных выходных данных или дезинформации, которая может серьёзно повлиять на безопасность пользователей, краже данных и другим формам вредоносных действий. В качестве примера исследователи предоставили следующий сценарий атаки:

Trail of Bits отмечает, что Sleepy Pickle может использоваться для поддержания скрытого доступа к системам ML, обходя средства обнаружения, так как модель компрометируется при загрузке файла Pickle в процессе Python.

Эта техника более эффективна, чем прямая загрузка вредоносной модели на Hugging Face, так как позволяет динамически изменять поведение модели или выходные данные без необходимости привлекать жертв к загрузке и запуску файлов. Миланов подчёркивает, что атака может распространяться на широкий круг целей, так как контроль над любым файлом Pickle в цепочке поставок достаточно для атаки на модели организации.

Sleepy Pickle демонстрирует, что продвинутые атаки на уровне моделей могут использовать слабые места в цепочке поставок, что подчёркивает необходимость

усиления мер безопасности при работе с моделями машинного обучения.