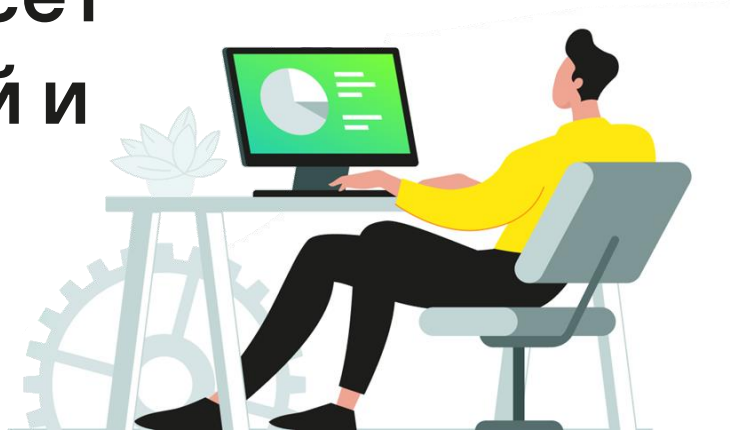
Опасности бесконтрольной разработки искусственного интеллекта: как их избежать

Александр Лискин,
Руководитель управления исследования угроз

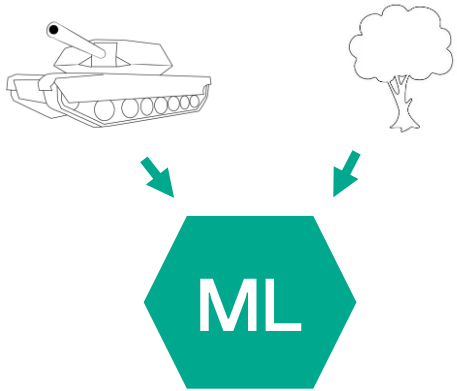
kaspersky

Системы машинного обучения проникают в нашу жизнь.
Их компрометация несет риски для организаций и индивидуальных пользователей.



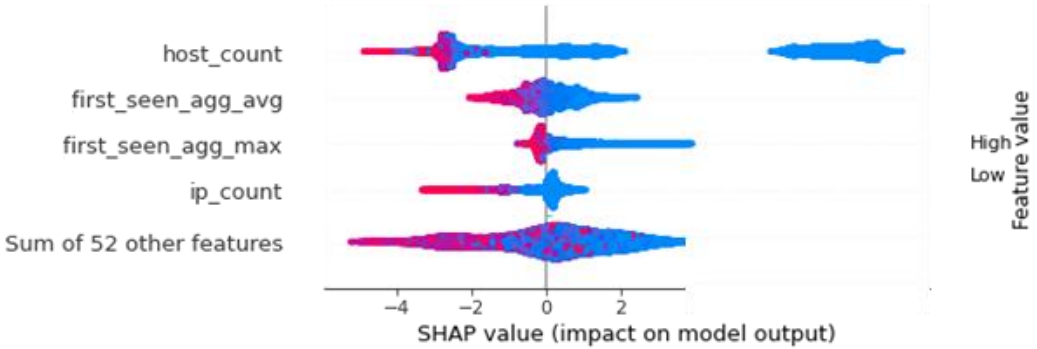
- Ошибки обучения.
- Атаки на алгоритмы.
- Классические проблемы безопасности.

- ML модели сложно анализировать.
- Тестовые выборки могут не выявить проблем.
- Проблема танка.



- Плохо работающая модель
  может нанести существенный ущерб.

- Наличие адекватных метрик качества.
  Правильное составление выборок:
  - Разделение выборки на train, test, validation;
  - Отложенные по времени выборки;
  - Выборки из альтернативного источника данных.
- Замеры качества работы модели после внедрения.
- Анализ: на какие признаки ориентируется модель.

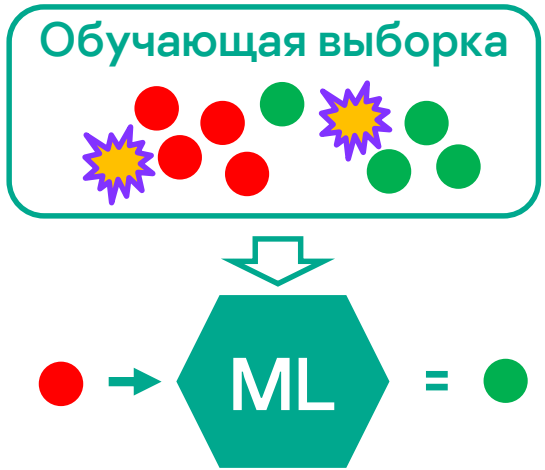


- Отравление обучающей выборки.
- Уклонение от обнаружения.
- Утечки данных из модели.
- ...

Атаки на отравление данных:

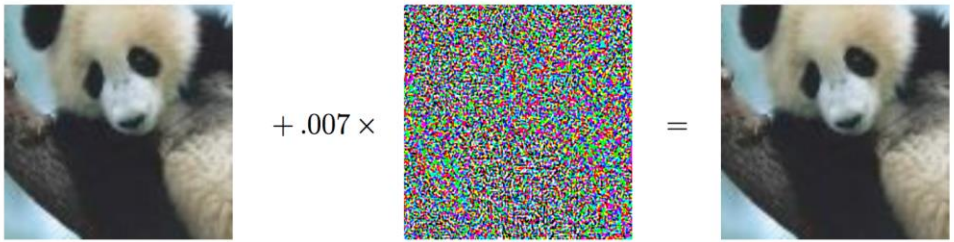
- *Model skewing* — загрязнение обучающей выборки с целью смещения у модели границы решения.

- *Backdoor attack* — внедрение в обучающую выборку примеров с определенными метками. Модель принимает неверное решение при возникновении данной метки.



- Возможности отравления данных в системе.
- Контроль и анализ входных данных.
- Сравнение модели с предыдущей стабильной версией.
- Контрольный набор данных.

Модель неправильно распознает специально сформированный объект.



«панда»

Уверенность
57.7%

шум

«гиббон»

Уверенность
99.3%

Такие атаки работают не только для изображений.

- Анализ применимости атак.
- Можно сделать модели устойчивее:
  - добавить в обучающую выборку атакующие примеры;
  - дистилляция модели;
  - использовать монотонные модели;
  - ...

- Модель может обладать приватной информацией.
- Есть техники извлечения данных из моделей.
- Например:
  - узнать персональные данные клиента;
  - узнать скрытые особенности модели.

- Анализ рисков.
- Не использовать персональные данные при обучении.
- Минимизировать запоминание частных случаев моделью.
- Предполагать возможность получения данных из модели.

- Закладки в открытом коде:
  - во фреймворках;
  - в моделях;
  - в данных.
- Классический взлом.
- Человеческий фактор.

- Лучшие практики защиты информации:
  - Контроль открытого кода, моделей и данных;
  - использование защитных решений.
  - ...

Спасибо!

Вопросы?

Александр Лискин

Руководитель управления исследования угроз

Alexander.Liskin@kaspersky.com

kaspersky