**Кейс: Распознавание эмоций в музыке для создания плейлистов настроения**

**Матрица источников и целей угроз ИИ-проекта**

|  |  |  |  |
| --- | --- | --- | --- |
| Источник угроз | Цели угроз | | |
| *Модели* | *Данные* | *Инфраструктура* |
| Киберпреступники | * Нарушения конфиденциальности: Кража модели * Нарушение целостности: Модификация модели * Управление моделью: Внедрение бэкдоров | * Утечка интеллектуальной собственности: Кража обучающих данных * Искажение рекомендаций: Манипуляция данными | * Отказ в обслуживании: DDoS-атака на сервис * Компрометация системы: Внедрение вредоносного ПО |
| Недовольные пользователи | * Блокировка работы модели: Попытки вывести модель из строя | * Снижение точности прогнозов: Некорректное использование обратной связи * Ухудшение качества модели: Массовая отправка некорректных данных | * Сбои в работе системы: Намеренная перегрузка сервиса |
| Конкуренты | * Потеря конкурентного преимущества: Копирование модели * Утечка технологий: Промышленный шпионаж | * Потеря конкурентного преимущества: кража обучающих данных | * Снижение производительности: Создание ложной нагрузки * Утечка данных: Попытка получить доступ |
| Обиженные сотрудники | * Утечка интеллектуальной собственности: Кража модели * Вывод модели из строя: внедрение логических бомб | * Потеря данных: Удаление/порча данных * Утечка персональных данных: Разглашение конфиденциальных данных | * Отказ в обслуживании: Блокировка доступа к серверу * Компрометация системы: Изменение настроек безопасности |

**Матрица рисков**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Вероятность** | **Влияние** | | | | |
| *Незначительное*  *< $10K* | *Минимальное*  *> $10K* | *Среднее*  *> $100K* | *Существенное*  *> $1M* | *Катастрофическое*  *> $10M* |
| **Почти наверняка**  **> 1 раза в год** |  |  | Некорректное использование обратной связи |  |  |
| **Вероятно**  **ежегодно** |  |  | Массовая отправка некорректных данных  Создание ложной нагрузки | DDoS-атака на сервис  Внедрение вредоносного ПО | Манипуляция данными |
| **Возможно**  **каждые 5 лет** | Намеренная перегрузка сервиса | Кража обучающих данных | Изменение настроек безопасности  Блокировка доступа к серверу | Кража модели  Кража обучающих данных  Разглашение конфиденциальных данных |  |
| **Маловероятно**  **каждые 20 лет** |  |  | Попытка получить доступ  Утечка данных |  | Внедрение логических бомб |
| **Редко**  **еще реже** |  |  |  | Промышленный шпионаж |  |

**Наиболее существенные риски**

1. Манипуляция данными
2. Внедрение вредоносного ПО
3. DDoS-атака на сервис
4. Кража обучающих данных
5. Кража модели
6. Разглашение конфиденциальных данных
7. Некорректное использование обратной связи

**Меры предотвращения/снижения риска**

1. Манипуляция данными:   
   Проверка и валидация входных данных, механизмы аудита данных.
2. Внедрение вредоносного ПО:   
   Регулярное обновление ПО, использование антивирусного ПО и IDS/IPS.
3. DDoS-атака на сервис:   
   Использование CDN, настройка брандмауэров и IDS/IPS.
4. Кража обучающих данных:   
   Шифрование данных, ограничение доступа.
5. Кража модели:  
   Шифрование модели, ограничение доступа.
6. Разглашение конфиденциальных данных:   
   Обучение сотрудников, применение соглашений о неразглашении (NDA).
7. Некорректное использование обратной связи  
   Валидация данных, вменяемые ограничения на количество обратной связи и её распределение

**Обновленная матрица рисков**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Вероятность | Влияние | | | | |
| Незначительное  < $10K | Минимальное  > $10K | Среднее  > $100K | Существенное  > $1M | Катастрофическое  > $10M |
| Почти наверняка  > 1 раза в год |  |  |  |  |  |
| Вероятно  ежегодно |  | Некорректное использование обратной связи |  |  |  |
| Возможно  каждые 5 лет | Намеренная перегрузка сервиса  Создание ложной нагрузки | Кража обучающих данных  Разглашение конфиденциальных данных  Массовая отправка некорректных данных | Изменение настроек безопасности  Блокировка доступа к серверу |  |  |
| Маловероятно  каждые 20 лет |  | DDoS-атака на сервис | Попытка получить доступ  Утечка данных | Манипуляция данными  Кража модели |  |
| Редко  еще реже |  |  | Внедрение вредоносного ПО  Кража обучающих данных | Промышленный шпионаж  Внедрение логических бомб |  |