АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

ЭКОНОМИКА
АНО «Цифровая экономика»

FIRST RUSSIAN DATA FORUM

Тренд объединения клиентских данных

Нейман Алексей, АБД

АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

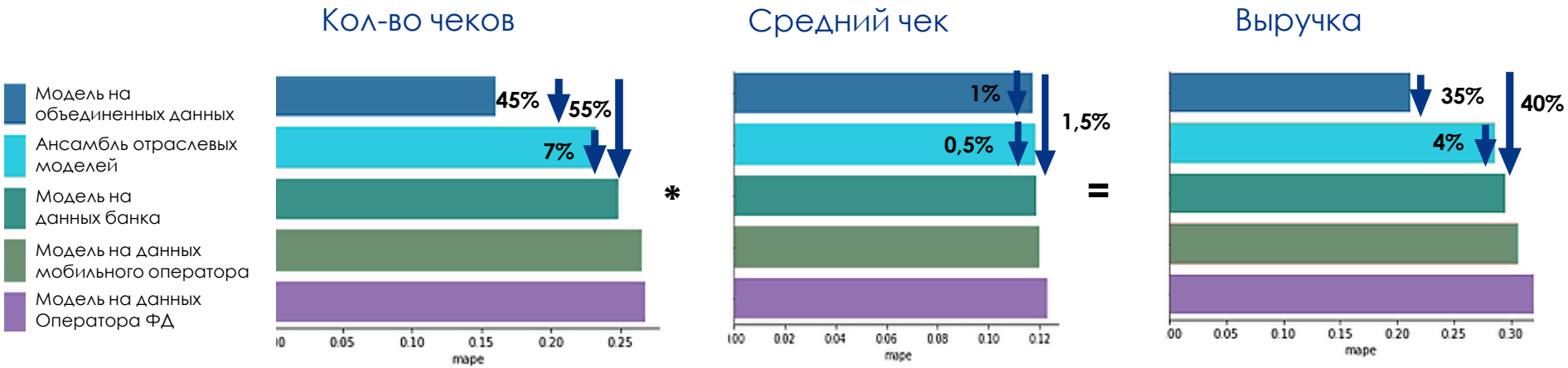
FIRST
RUSSIAN
DATA
FORUM

Есть ли потенциал в объединении клиентских данных из разных источников ?

Эффект моделирования на объединенных данных

FIRST RUSSIAN DATA FORUM

Эксперимент **АБД «Собственное дело»** по моделированию выручки торговой точки на данных банка, ОМС, ОФД, ГИС



Результаты **моделирования целевой функции на объединенных данных** значимо лучше: на 35-45% чем у ансамбля и на **40-55% лучше**, чем на частных данных поставщиков

АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

FIRST
RUSSIAN
DATA
FORUM

Каковы риски объединения клиентских данных?

Риск-модель обработки клиентских данных

FIRST RUSSIAN DATA FORUM

ТЕХНОЛОГИИ ЗАЩИТЫ ПЕРСОНАЛЬНЫХ ДАННЫХ

1 ОРГАНИЗАЦИОННЫЕ И ОПЕРАТИВНО-ТЕХНИЧЕСКИЕ МЕРЫ

Организационные и оперативно технические меры ведут к управлению контекстными рисками, которые оцениваются с помощью скоринг-модели:

$$P_{\text{контекстные риски}} = \frac{\sum_{j=1}^{n} \omega_j K_j}{\sum_{j=1}^{n} \omega_j}$$

Контекстные риски отражают способность организаций противостоять угрозам утечек данных и поддерживать процессы управления персональными данными, снижая общие угрозы

2 КРИПТОГРАФИЧЕСКИЕ МЕТОДЫ

3 МЕТОДЫ ПСЕВДОНИМИЗАЦИИ

Методы псевдонимизации – методы защиты персональных данных, при которых прямые и/или косвенные атрибуты в конкретных наборах данных заменяются на один или несколько искусственных идентификаторов

$$P_{\text{риски данных}} = \frac{k}{\langle r|R \rangle}$$

При расчете рисков данных в условиях применения методов псевдонимизации величина рисков обратна пропорциональна затраченным ресурсам (например, времени на взлом при атаках прямым перебором)

4 ОБЕЗЛИЧИВАНИЕ (АНОНИМИЗАЦИЯ)

Методы обезличивания нацелены на исключение связи отдельных атрибутов с идентификаторами. Основная идея для данного класса методов: выделении групп схожих записей (классов эквивалентности), которые могут быть отнесены более чем к одному физическому лицу.
В этом случае нарушается основное свойство определения персональных данных – соотнесение информации с прямо *или косвенно определенным или определяемом* физическом лицом.

$$P_{\text{риски данных}} = \frac{k}{\langle \check{E} \rangle}$$

Вероятность повторной идентификации обратна пропорциональна размеру наименьшего класса эквивалентности в наборе

5 КОНФИДЕНЦИАЛЬНЫЕ ВЫЧИСЛЕНИЯ

В рамках таких методах используются комплексные криптографические схемы, а в качестве метрик риска – оценки вероятности повторения значений совокупности ключей и результатов вычисления крипто-функций на всем пространстве определения (так называемая крипто-игра).

$$P = \frac{r_{train} - r_{control}}{1 - r_{control}}$$

Доверительный интервал биноминального распределения по методу Уилсона

6 СТАТИСТИЧЕСКИЕ МЕТОДЫ И МАШИННОЕ ОБУЧЕНИЕ

Данная группа методов основана применении методов машинного обучения или статистического подхода. Методы объединяет идея замены исходных наборов данных новыми информационными массивами, сохраняющими свойства исходных наборов, но с измененными значениями.

$$P_{\text{риски данных}} = \max\left(\frac{1}{N}\sum_{s=1}^{n}\left(\frac{1}{f_s} \times I_s\right), \frac{1}{N}\sum_{s=1}^{n}\left(\frac{1}{F_s} \times I_s\right)\right)$$

Обобщенная формула Эль-Эмама

Для случаев применение дифференциальной приватности количество шума, добавленного к данным, может быть количественно определено с помощью значения ε (меньшее значение ε подразумевает более высокий уровень защиты).

Некоторые из указанных методов могут применяться последовательно, что позволяет дополнительно факторизовать (разложить на множители формулу расчета, например:

$$P_{\text{повторной идентификации}} = P_{\text{контекстные риски}} \times (P_{\text{синтетические наборы}} \times P_{\text{обезличивание}})$$

Измерение риска по 3-м бизнес-кейсам

FIRST RUSSIAN DATA FORUM

№	КЕЙС	СЦЕНАРИЙ ОБМЕНА ИНФОРМАЦИЕЙ	МЕТОДЫ ЗАЩИТЫ	ОПИСАНИЕ МЕТОДОВ	СЛОЖНОСТЬ АТАКИ	РИСК РЕИДЕНТИФИ-КАЦИИ	ПРИМЕЧАНИЕ
1	**СКОРИНГ С ПРИВЛЕЧЕНИЕМ 3-ИХ СТОРОН** ОБМЕН ДАННЫМИ С ИСПОЛЬЗОВАНИЕМ ОДНОСТОРОННИХ КРИПТОГРАФИЧЕСКИХ ФУНКЦИЙ	Сторона А (финансовая организация) формирует массив своих клиентов (представленных мобильными номерами телефонов) и запрашивает сторону В (скоринговое агентство) для формирования скоринга на основании обогащенной информации. сторона В формирует массив данных с номерами телефонов и соответствующих им скоринговых коэффициентов, возвращает его стороне А. Требуется защита схемы обмена.	Хэширование MD5 с использованием статической соли	Нас стороне А номера телефонов хэшируется с заранее согласованной солью – случайной строкой	НИЗКАЯ 1 мин – 10 минут	0.3-0.7 (в зависимости от соли)	На расшифровку идентификатора понадобится 10 минут обычного ноутбука
			Использование более сильных функций хеширования (SHA-3/Kessack)	Дополнительно первый метод усиливается использованием более сложной хэш функцией с двойным хешированием	СРЕДНЯЯ 3 суток-1 месяц	0.1	Для расшифровки идентификатора понадобится от трех суток до месяца при работе на мощном персональном компьютере без распараллеливания
			Использование дополнительно динамической соли	Дополнительно организуется отдельный сеанс с передачей набора динамических солей для каждого номера	ВЫСОКАЯ 3 месяца – 3 года	0.00001	Для расшифровки идентификаторов понадобится от 3 месяцев до 10 лет при использовании компьютерного кластера до 4-х узлов
2	**МАРКЕТИНГОВОЕ КАСАНИЕ** НАХОЖДЕНИЕ ПЕРЕСЕЧЕНИЯ ПРИВАТНЫХ НАБОРОВ ДЛЯ ПОСТРОЕНИЯ АНАЛИТИЧЕСКИХ МОДЕЛЕЙ	Сторона А (маркетинговая платформа) формирует данные о просмотрах рекламы пользователями, представленными номерами телефонов. Сторона В (финансовая организация) представляет информацию о платежах по рекламируемому продукту, а также основной профиль ФЛ, осуществившего платеж. Сторона С совмещает наборы информации от А и В, строит модель для машинного обучения, позволяющую осуществить маркетинговое таргетирование. Требуется защита схемы обмена.	Формирование для телефонов микросегментов и классическое обезличивание	Представление телефонов, как чисел из заданного промежутка и затем обобщение информации для выбранного диапазона	НИЗКАЯ	0.5	Для сопоставления данных с ФЛ потребуется расчет на компьютере средней мощности в течении 10-18 часов
			Использование синтетических наборов на каждом этапе	Формирование синтетических наборов для заданных параметров (векторов) и затем построение обобщенной модели	СРЕДНЯЯ	0.08	Для сопоставления данных с ФЛ потребуется дополнительные наборы данных (утечки) и не менее 3 недель обработки на компьютерах средней мощности
			Реализация PS^3I схемы многосторонних безопасных вычислений	Сложная схема обмена информации на основании нескольких ключей и гомоморфного шифрования	ВЫСОКАЯ	0.00000001	Для сопоставления записей с ФЛ потребуется дополнительная информация и соответствующая обработка на компьютерах не менее 10 лет
3	**ОЗЕРО ДАННЫХ** ОБЕЗЛИЧИВАНИЕ ДЛЯ БОЛЬШИХ ИНФОРМАЦИОННЫХ КОМПЛЕКСОВ	СТОРОНА А (финансовая организация) реализует хранилище /озеро данных для управления данными своими клиентов (цифровые профили, майнинг процессов, маркетинг). К озеру данных имеют доступ аналитики данных, а также привлеченные третьи стороны (строят модели и прогнозы). Необходима модель защиты в условиях постоянно наращиваемых больших данных	Маскеризация данных	Замена данных или их частей символами-заменителями	СРЕДНЯЯ (в зависимости от степени обобщения)	0.2-0.5	Для построения соответствия записей с ФЛ потребуется дополнительная информация и обработка в течении 7 дней на компьютере средней мощности
			Итерационная схема с использованием расчета метрик обезличивания	Формирование схемы с обобщениями и подавлениями редких записей, на каждом этапе проведение измерений метрик k-anonymity	СРЕДНЯЯ/ ВЫСОКАЯ Устойчивость к атакам связывания	0.2	Для построения соответствия записей с ФЛ потребуются дополнительные сведения и обработка не менее 80 дней на компьютере средней мощности
			Многомерная схема обезличивания на основе метода Мондриана	Разбиение исходных наборов на зоны с одинаковыми классами эквивалентности и проведение («жадного») поиска	ВЫСОКАЯ	< 0.1	Для построения соответствия между записями и ФЛ потребуются дополнительные сведения не менее 2 лет на компьютере большой мощности

Подход к управлению рисками объединения данных

FIRST RUSSIAN DATA FORUM

Риски обработки клиентских (обезличенных) данных могут (и должны) быть измерены для конкретного бизнес-кейса

Существуют техники и технологии снижения риска реидентификации до околонулевых значений

Обработка клиентских данных с околонулевыми рисками реидентификации может считаться обработкой обезличенных данных

Разработка цифровых продуктов (моделей, сервисов, и пр.) на объединенных обезличенных данных при рисках реидентификации ниже пороговых значений позволяет раскрыть потенциал имеющихся данных

АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

FIRST
RUSSIAN
DATA
FORUM

А нужны ли реальные клиентские данные?

Синтетические данные пригодны для решения задач

FIRST RUSSIAN DATA FORUM

Проведены эксперименты, которые включали в себя следующие архитектуры нейронных сетей: RNN, VAE, CVAE, Transformer, PAR.

Результаты:

Лучшие результаты показали эксперименты основанные на архитектуре Transformer

	Возраст rmse			Пол roc-auc		
Модель	train	val	test	train	val	test
Transformer	**7.15**	**9.56**	10.21	**0.94**	0.71	0.67
Transformer параллельные ветки	11.06	11.13	**10.18**	0.73	**0.72**	**0.71**
CVAE	**7.15**	9.63	10.72	**0.94**	**0.72**	0.66

Параметры качества

Качество метода генерации синтетики определяет значение метрик и близость значений, полученных на реальных данных к значениям метрик, полученных на синтетических данных.
В данном случае, метриками выступали ROC-AUC (метрика качества бинарной классификации) и RMSE (метрика качества для регрессии)

Использование моделей, обученных на синтетике, для решения задач на реальных данных видится перспективным направлением, требующем дополнительной проработки

АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

FIRST
RUSSIAN
DATA
FORUM

Спасибо за внимание!