

СИНТЕЗ ДАННЫХ ДЛЯ КРЕДИТНОГО СКОРИНГА

Загibalов М.А.

Научный руководитель: Брагин А.Д.

Томский политехнический университет, ИШИТР, студент гр.8К43

e-mail: maz22@tpu.ru

Аннотация

В данной работе представлен анализ перечня персональных данных, применяемых в кредитном скоринге. Анализ включает в себя рассмотренные правовые ограничения, ключевые параметры оценки и подход к их синтезу в рамках модели, позволяющий сохранить статистическую ценность данных, сохранив конфиденциальность.

Ключевые слова: персональные данные, кредитный скоринг, модель обучения.

Введение

В ряде задач, решаемых с помощью машинного обучения, персональные данные играют важную роль. Примером такой задачи является кредитный скоринг. Информация о возрасте, доходе, семейном положении и других характеристиках заемщика позволяет строить более точные модели оценки кредитоспособности. Однако использование таких данных сопряжено с рядом правовых и этических ограничений. Законы о защите персональных данных не допускают свободное распространение и использование информации, позволяющей идентифицировать конкретное лицо.

Создание искусственных наборов данных, сохраняющих статистические свойства оригинальных, но не содержащих реальных персональных записей, является одним из способов решения данной проблемы. Такой подход позволяет разрабатывать и тестировать модели машинного обучения без риска нарушения конфиденциальности.

Персональные данные

В Российской Федерации существует определение персональных данных (ПД), которое содержится в Федеральном законе № 152-ФЗ [1]. Согласно статье 3 этого закона, персональными данными считается любая информация, которая прямо или косвенно относится к конкретному физическому лицу. К персональным данным относятся, например, ФИО, дата рождения, адрес проживания или регистрации, семейное положение, образование, профессия, пол, гражданство, сведения о документе, удостоверяющем личность (например, серия и номер паспорта), контактные данные (электронная почта, номер телефона), номера счетов или карт, а также другие финансовые реквизиты [2]. Идентификационные номера, такие как ИНН (идентификационный номер налогоплательщика) и СНИЛС, также считаются персональными данными.

Проблемы использования персональных данных

Обеспечение защиты персональных данных накладывает ряд ограничений в исследованиях в сфере кредитного скоринга. Финансовые организации накапливают большие объёмы информации о своих клиентах, но не могут свободно обмениваться этими данными или использовать их в исследованиях из-за требований законодательства о конфиденциальности. Это ограничивает доступ к реальным данным, необходимым для разработки и тестирования моделей оценки кредитного риска, что замедляет прогресс в этой области.

В соответствии с Федеральным законом №152-ФЗ, регулирующим обработку персональных данных, операторы данных, обрабатывающие информацию о физических лицах, обязаны соблюдать ряд требований. В частности, они должны получить согласие

субъекта на обработку его персональных данных, обеспечить их безопасность, не собирать избыточные данные и хранить их не дольше, чем это необходимо для целей обработки.

Банки обычно получают персональные данные на основании договора с клиентом, например, заявления на кредит. Однако передача персональных данных третьим лицам или их публикация требует отдельного основания. Как правило, публикация реальных данных клиентов, даже в обезличенном виде, запрещена.

Альтернатива реальным данным

Генерация искусственных данных, статистически схожих с исходными, является одним из путей решения данной проблемы. Синтетические данные позволяют обучать модели без использования реальных персональных сведений, что позволяет избежать рисков для приватности и соблюдать требования законодательства о защите данных [3, 4]. Например, вместо использования базы реальных клиентов банка для обучения модели скоринга, можно сгенерировать «выдуманные» записи клиентов, которые сохраняют статистические закономерности оригинала, но не содержат личной информации конкретных людей [5, 6].

Персональные данные в задачах скоринга

Модели кредитного скоринга учитывают широкий спектр характеристик и сведений о пользователе. Данные о заемщике, которые можно считать персональными, представлены в таблице 1.

Таблица 1. Пример подхода к синтезу персональных данных заемщика

ПД	Назначение при синтезе	Подход к синтезу данных
ФИО	Используются для идентификации пользователя, не влияет на конечную оценку	Не следует использовать реальные данные. Вместо этого можно применить псевдонимы.
Контакты	Используется для только для связи.	В модели данные не участвуют, в связи с этим их можно опустить.
Паспортные данные	Используются только для верификации.	Нахождение таких данных о реальном человеке в наборе строго запрещено. Следует их исключить.
ИНН и СНИЛС	Используются только для идентификации.	Эти данные не используются в модели скоринга, поэтому следует их исключить.
Пол	Некоторые модели учитывают, но это не всегда важно.	Пол относится к общим ПД, его указание возможно только с согласия. Для модели возможен вариант распределить данные равномерно (50/50) между мужчинами и женщинами.
Семейное положение	Может оказывать влияние на риски.	Распределять заемщиков по группам с выдержанным основным признаком (напр. женат/холост/разведен)
Трудовой стаж, должность	Используется в скоринге, поскольку оказывает влияние на платежеспособность.	Можно отнести к определенной категории (трудоустроенный, безработный, индивидуальный предприниматель)
Доход	Ключевой показатель при кредитном скоринге.	Доходы являются ПД, поэтому их публикация строго запрещена. В синтетическом наборе можно сгенерировать данные, сохраняющие основную суть оригинала.
Адрес проживания (регистрации)	Может использоваться на уровне города или региона.	Данные сведения стоит обобщить, поскольку использование точных данных раскрывает личность и веса в модели не имеют. Лучше будет обобщить данные, сохранив город или регион.
Возраст	Часто используется для скоринга.	Точная дата рождения раскрывает личность. Допустимо рассчитать возраст заемщика и отнести его к соответствующей возрастной категории.
Кредитная история	Ключевой показатель при кредитном скоринге.	Данные кредитной истории являются приватными, но их агрегированные метрики важны для модели. При синтезе необходимо сохранить корреляции между ними.

Важно понимать, что при публикации обезличенного датасета, сгенерированные данные не должны позволять обратной идентификации личности.

Можно заметить, что часть данных не влияет на расчет кредитного скоринга, поэтому нет смысла включать их в обучающий набор. В данных, используемых для обучения модели, должны быть указаны следующие параметры: возраст, пол, семейное положение, доход, стаж, информация о предыдущих кредитах. При обезличивании данных необходимо синтезировать эти ключевые атрибуты с сохранением статистических закономерностей, чтобы модели, обученные на синтетических данных, давали результаты, близкие к результатам обучения на реальных данных.

Заключение

В рамках данной работы был проведен анализ ключевых аспектов использования персональных данных в задачах кредитного скоринга. Основное внимание было уделено проблеме ограниченного доступа к реальным данным из-за требований конфиденциальности, а также способу ее решения – генерации синтетических данных. Такой метод позволяет сохранить статистические закономерности исходных данных, при этом исключая риски нарушения приватности. Специально для этого был сформирован перечень персональных данных, необходимых для корректной работы модели кредитного скоринга.

Использование синтетических данных открывает новые возможности для разработки и тестирования моделей машинного обучения без нарушения законодательства о защите персональной информации. Это особенно актуально для финансового сектора, где соблюдение конфиденциальности клиентов является критически важным.

Список использованной литературы

1. Федеральный закон РФ от 27.07.2006 № 152-ФЗ «О персональных данных» // Собрание законодательства РФ. – 2006. – № 31 (ч. 1). – Ст. 3451.
2. Skillfactory. Что такое персональные данные и как с ними работать // Skillfactory: блог. – [Электронный ресурс]. – URL: blog.skillfactory.ru/chto-takoe-personalnye-dannye-i-kak-s-nimi-rabotat/ (дата обращения: 01.04.2025).
3. Skillfactory. Синтетические данные в 2025 // Skillfactory: блог. – [Электронный ресурс]. – URL: blog.skillfactory.ru/sinteticheskie-dannye/ (дата обращения: 01.04.2025).
4. Synthetic data – history and usage in privacy context. // Wikipedia. – [Электронный ресурс]. – URL: en.wikipedia.org/wiki/Synthetic_data (дата обращения: 01.04.2025).
5. Muñoz-Cancino R., Bravo C., Ríos S.A., Graña M. Assessment of creditworthiness models privacy-preserving training with synthetic data // International Conference on Hybrid Artificial Intelligence Systems. – Cham: Springer International Publishing, 2022. – P. 375–384.
6. Financial Conduct Authority. Using Synthetic Data in Financial Services. – [Электронный ресурс]. – URL: fca.org.uk/publications/corporate-documents/report-using-synthetic-data-financial-services (дата обращения: 01.04.2025).