

Использование машинного обучения для выявления случаев мошенничества на уровне транзакций

Антипова Евгения Сергеевна
РЭУ им. Г.В. Плеханова
Студент

Кушнерева Екатерина Владиславовна
РЭУ им. Г.В. Плеханова
Студент

Аннотация

В статье рассматриваются недостатки экспертных систем и причины использования машинного обучения для обнаружения мошеннических транзакций. Помимо этого, исследованы методы машинного обучения, такие как логистическая регрессия, дерево решений, случайный лес и нейронные сети.

Ключевые слова: машинное обучение, мошенничество, информационные технологии, информационная безопасность.

Using machine learning to detect fraud at the transaction level

Antipova Evgenia Sergeevna
Plekhanov Russian University of Economics
Student

Kushnereva Ekaterina Vladislavovna
Plekhanov Russian University of Economics
Student

Abstract

The article considers the shortcomings of expert systems and the reasons for using machine learning to detect fraudulent transactions. In addition, machine learning methods such as logistic regression, decision tree, random forest, and neural networks have been studied.

Keywords: machine learning, fraud, information technology, information security.

В наше время объемы транзакций растут, появляются новые платежные инструменты, а мошеннические схемы лишь усложняются. В связи с этим вычислять мошенников становится все сложнее, а финансовые потери только растут. Эти факторы привели к тому, что для выявления случаев мошенничества стали использовать машинное обучение.

Для обнаружения мошеннических транзакций широко используются экспертные системы, которые работают благодаря логическим выражениям и набором правил, написанных экспертами. Данные системы действуют следующим образом: каждая транзакция выполняет ряд условий, и некоторые из них повышают оценку риска. Если эта оценка перейдет определенный порог, то транзакцию отправляют на дополнительную проверку. Такие системы выявления мошенничества имеют определенные недостатки:

1. Они субъективны, так как полностью зависят от эксперта;
2. Не могут выявить новые мошеннические схемы;
3. Имеют большое количество ложных срабатываний;
4. Могут обнаружить только очевидные корреляции.

Экспертные системы нуждаются в актуализации и масштабировании, должны учитывать большее количество параметров, оценки веса риска должны определяться не экспертами, так как те не могут быть до конца объективными. Все факторы, описанные выше, привели к тому, что в банковской сфере для обнаружения случаев мошенничества стали использовать машинное обучение.

Процесс обнаружения мошенничества с использованием машинного обучения начинается со сбора и сегментации данных. Затем модель машинного обучения работает с обучающими наборами для прогнозирования вероятности мошенничества.

Извлечение данных

Как правило, данные будут разделены на три разных сегмента - обучение, тестирование и перекрестная проверка. Алгоритм будет настраиваться по частичному набору данных и параметров, полученных на тестовом наборе. Полезность данных измеряется с использованием набора кросс-валидации. Высокопроизводительные модели будут затем проверены на различные случайные расщепления данных, чтобы обеспечить согласованность результатов.

Предоставление учебных комплектов

Основное применение машинного обучения, используемого при обнаружении мошенничества - это предсказание. Мы хотим предсказать значение некоторого результата (в данном случае логическое значение, которое истинно, если платеж является мошенническим и ложным в противном случае), учитывая некоторые входные значения. Данные, которые используются для обучения моделей machine learning, состоят из записей с выходными значениями для различных входных значений. Записи часто берутся из исторических данных.

Модели знаний

Модели знаний являются важным шагом в прогнозировании мошенничества или аномалии в наборах данных. Определяется, как сделать это предсказание на основе предыдущих примеров входных и выходных данных. Можно еще разделить проблему прогнозирования на два типа задач: классификация и регрессия. Ниже рассмотрены некоторые из таких задач.

1. Логистическая регрессия

Регрессионный анализ - популярный, давний статистический метод, который измеряет силу причинно-следственных связей в структурированных наборах данных. Регрессионный анализ, как правило, становится более сложным при применении к обнаружению мошенничества из-за количества переменных и размера наборов данных. Он может обеспечить ценность путем оценки прогностической способности отдельных переменных или комбинаций переменных в рамках более масштабной стратегии мошенничества. В этих методах аутентичные транзакции сравниваются с мошенническими для создания алгоритма. Эта модель будет определять, является ли новая транзакция мошеннической или нет. Для очень крупных продавцов эти модели специфичны для их клиентской базы, но обычно применяются общие алгоритмы.

2. Дерево решений

Это зрелое семейство алгоритмов машинного обучения используется при автоматизации создания правил для задач классификации. Алгоритмы дерева принятия решений могут использоваться для задач моделирования прогнозирования классификации или регрессии. Это набор правил, которые получаются с использованием примеров мошенничества, с которыми сталкиваются клиенты. Создание дерева игнорирует несущественные функции и не требует значительной нормализации данных. Дерево можно проверить, и мы можем понять, почему было принято решение, следуя списку правил, инициированных определенным клиентом. Это дает оценку вероятности мошенничества на основе более ранних сценариев.

3. Случайный лес

Техника Random Forest использует комбинацию нескольких деревьев решений для повышения эффективности классификации или регрессии. Это позволяет избежать ошибки, которая может существовать в одном дереве, а также повышает общую производительность и точность модели, сохраняя при этом способность интерпретировать результаты и предоставлять объясняемые оценки пользователям. Время работы данной техники довольно быстрое, и она может справиться с несбалансированными и отсутствующими данными. Недостаток заключается в том, что при использовании для регрессии эта техника не может предсказывать за пределами диапазона данных обучения.

4. Нейронные сети

Это отличное дополнение к другим методам, и оно улучшается с использованием данных. Нейронная сеть является частью технологии когнитивных вычислений, где машина имитирует, как работает человеческий мозг и как он наблюдает за узорами. Нейронные сети полностью адаптивны, так как способны учиться на образцах поведения. Они могут адаптироваться к изменениям поведения обычных транзакций и определять модели транзакций мошенничества. Процесс нейронных сетей чрезвычайно быстрый и может принимать решения в режиме реального времени.

Нужно заметить, что несмотря на то что введение в работу машинного обучения связано с высокими инфраструктурными издержками, строгими правилами и риском замены существующих технологий, этот метод невероятно эффективен и многие банки вводят его уже сейчас.

Библиографический список

1. Вьюгин В.В. Математические основы машинного обучения и прогнозирования. М., 2013.
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. 2003.
3. Rashid T. Make Your Own Neural Network. 2016.
4. Фридман А.Я. Ситуационное моделирование иерархических динамических систем // Информационные и математические технологии в науке и управлении. 2018. №1 (9). С. 5-15.
5. Жуков А.В. Модификация алгоритма случайного леса для классификации нестационарных потоковых данных. / А.В. Жуков, Д.Н. Сидоров // Вестник Южно-Уральского государственного университета. Серия: Математическое моделирование и программирование. 2016. №9 (4). С. 86–95.