1 слайд

Всем добрый день!
Темой моей дипломной работы является «Применение графов для классификации финансовых транзакций»

2 слайд

Задачей данной работы является анализ финансовых транзакций с целью выявления нетипичных или мошеннических операций. Ее можно разбить на две подзадачи:

Первая - Нужно реализовать классификатор, который будет способен отличить мошеннические транзакции от не мошеннических

Вторая – на основе транзакций построить графы финансовых переводов. На них спроектировать новые признаки, вроде расстояния до ближайшего мошенника, количество мошенников в круге радиуса n и т.п. Добавить полученные признаки в модель и проверить их значимость

3 слайд

Источником данных является сайт онлайн сообщества специалистов по анализу данных и машинному обучению kaggle.com

Данные синтетически сгенерированы программой PaySim – выборка моделируется на основе реальных финансовых операций сервиса мобильных переводов компании, которая предоставляет свои услуги в 14 африканских странах. При генерации выборки намеренно симулируется мошенническое поведение, чтобы проверить эффективность тех или иных методов борьбы с ним

Этот набор создан исключительно для kaggle симулирует один месяц работы сервиса, содержит порядка шести с половиной миллионов транзакций, 8213 из которых мошеннические.

4 слайд

Исходный набор данных состоит из 10 признаков. *Кратко перечисляю каждый.* isFraud является целевым признаком, который будет предсказываться.

5 слайд

Был проведен предварительный анализ данных, на котором выявлены некоторые вещи, лежащие на поверхности: *перечисляются данные из таблиц. Нужно упомянуть о возможной разреженности графов.*

Типичной для мошеннических транзакций является ситуация, представленная на иллюстрации: есть клиент с большим количеством приемов, один из которых мошеннический.

6 слайд

Перейдем к построению модели классификации. Задачей классификации является построение модели, которая будет способна отнести отдельный исследуемый объект к тому или иному классу. В данной работе используются модели классификации с учителем, то есть сначала выборка бьется на обучающую и тестовую. Затем происходит обучение, а затем тест. Для сравнения будут использованы три метода:

Случайный лес, метод к ближайших соседей и наивный байесовский классификатор.

7 слайд

Чтобы было с чем сравнивать, было проведено тестирование модели на исходных сырых данных.

Для анализа результатов используются матрица ошибок и такие метрики как

присижн (точность), которая показывает какую долю положительных объектов, действительно являющихся таковыми определил алгортим.

Реколл (полнота) которая показывает какую долю положительных объектов из всех положительных объектов определил алгоритм.

И Ф мера, которая является средним гармоническим между полнотой и точностью

Случайный лес проявил себя лучше остальных, k ближайших соседей тоже показал неплохой результат, а вот наивный байес оказался довольно плох.

8 слайд

После предварительного анализа и тестирования модели на сырых данных была произведена процедура проектирования признаков. То есть на основе исходных данных создаются вспомогательные, которые предположительно должны улучшить результат работы классификатора.

Для начала, чтобы немного сбалансировать выборку, транзакции были просеяны до порядка 450 тысяч, мошеннические сохранены в полном объеме.

Затем некоторые признаки убраны за ненадобностью: *перечисление убранных признаков.*

Шаг заменяется 24 часовым эквивалентом. *Перечисляются добавленные признаки.*

9 слайд

Далее была произведена процедура отбора признаков. То есть выявление наиболее значимых и отброс менее значимых признаков. На слайде можно видеть коэффициенты значимости. Отбор признаков происходил с помощью метода рекурсивного отсечения признаков, суть которого заключается в том, что модель обучается на исходном наборе признаков, затем итеративно отбрасывает наименее значимые, пока не останется n наиболее значимых.

Данная процедура способствует улучшению качества обучения модели. В результате были оставлены признаки *их перечисление.*

10 слайд

Сравним результат работы модели после проектирования отбора признаков. Случайных лес и метод ближайших соседей улучшили свои показатели, метод ближайших соседей даже вырвался вперед. А вот наивный байес стал только хуже.

11 слайд

Чтобы еще попытаться улучшить результаты, была произведена процедура нормализации данных. То есть преобразование всех признаков в числовой диапазон от нуля до единицы. Был использован минимаксный метод.

12 слайд

После нормализации результаты случайного леса и метода к ближайших соседей стали еще лучше, случайный лес вообще приблизился к идеалу. Наивный байес предсказал все с точностью да наоборот.

Результаты, показанные случайным лесом очень высоки, что наводит на мысли о неправдоподобности результатов. Скорее всего такие результаты являются следствием синтетической природы выборки.

13 слайд

Возвращаясь к постановке задачи, второй частью было проектирование признаков на основе графов. Но реализовав прикладную программу, визуализирующую графы (*мб рассказать о интерфейсе программы),* стало ясно, что из этого мало что выйдет: общий граф разбит на кучу маленьких обособленных подграфов, максимальная цепочка на всем исходном наборе – 9 клиентов. Все остальные примерно 2-4.

Таким образом расстояние до ближайшего мошенника практически всегда либо отсутствует, либо равно 0 или 1.

Количество мошенников в круге радиуса n отслеживать нет смысла, потому что почти нет достаточно длинных цепочек переводов.

14 слайд

В результате выполнения работы: *на слайде*

Спасибо за внимание!