Домашнее задание №5: Антифрод: подготовка данных, кластеризация и МЛ

Это домашнее задание можно доделывать: со штрафом 30%, если исправления присылаются не позже 24.12 (и тогда ещё можно получить "хорошо" или "отлично"), или со штрафом 50% в более позднее время (но тогда уже только на "зачёт").

Задание 1 (5 баллов) - первичная обработка данных

Возьмите файл dataset_16. Проведите первичную обработку данных: с помощью изучения распределений, корреляционного анализа и метода главных компонент определите столбцы, не интересные для анализа (скоррелированные с другими столбцами или не вносящие весомый вклад в общую дисперсию).

Для сильно скоррелированных пар столбцов подберите вид зависимости одного столбца от другого (если получится).

В качестве ответа на задание ожидается ipynb с кодом, пояснениями и графиками, из которых можно понять, каким образом и почему вы рекомендуете исключить из анализа те или иные столбцы, а также уравнения найденных зависимостей.

Датасет содержит специально заготовленные "ловушки", сходные тому, что встречается в реальных данных.

Задание 2 (7 баллов) - поиск аномалий, кластеризация

Возьмите файл dataset_16_2. Каждая строка представляет из себя факторы, описывающие действия пользователя на одном из сервисов Яндекса за какой-то день. Часть пользователей являются фродовыми.

Проведите первичную обработку данных аналогично заданию 1: определите и исключите столбцы, не интересные для анализа.

Найдите строки, соответствующие аномальным (в каком-либо смысле) пользователям. Можно использовать определение аномальности из питон-ноутбуков с лекции (по попаданию в крайние квантили распределений), можно предложить свое.

Найдите кластера аномальных пользователей с помощью k-средних.

Найдите переменные, которые похожи на категориальные (или могут быть интерпретированы как категориальные). Изучите распределение аномальности в кластерах, образованных значениями этих переменных.

Сделайте регрессию аномальности по имеющимся в файле столбцам (любым известным вам способом). Оцените точность, полноту.

В качестве ответа на задание ожидается ipynb с ходом решения задачи, и вердикты по каждому пункту:

(1) - исключаем такие-то столбцы,

(2) - такие-то строки являются аномальными по выбранной методике,

(3), (4) - такие-то кластера включают аномальных пользователей (либо кластеризация не работает)

(5) - подобранная регрессия, точность и полнота.

Задание 3 (8 баллов) - Машинное обучение для поиска фродовых пользователей

В архиве dataset163.tar.gz содержится два файла: dataset_16_3_learn и dataset_16_3_test.

Файл learn содержит 140 столбцов и порядка 9к строк, каждая строка представляет из себя факторы, описывающие действия пользователя на одном из сервисов Яндекса за какой-то день.

Столбцы 0 и 1 - это хеши от ip-адреса пользователя: хеш от полного ip-адреса (hash(192.168.255.255)) и хеш от первых трех октетов (hash(192.168.255))

Столбец 2 - это флаг is_fraud_user, равный 1 для фродовых пользователей и 0 для хороших.

Последующие столбцы - это какие-то факторы, характеризующие сессию пользователя. О значении некоторых из них можно догадаться по их виду.

Файл test аналогичен файлу learn, но не содержит столбца с флагом is_fraud_user.

Выполните следующие действия:

1. проведите анализ и первичную обработку данных.
2. Сделайте машинно-обученный классификатор, который зафитит флаг is_fraud_user в файле learn. Рассчитайте точность, полноту, f1-меру. Попробуйте несколько различных классификаторов, выберите среди них лучший.

3. Проведите манипуляции над столбцами, которые позволят повысить точность классификатора. Определите, на что лучше заменять отсутствующие значения (missing values).
4. Постарайтесь использовать в машинном обучении данные об ip-адресе пользователя следующим образом:

- выберите один из ip-хешей, сагрегируйте значения факторов, для каждого ip-хеша посчитав среднее значение факторов
- проведите регрессию, подобрав mean(is_fraud_user) на основе mean(factor_1) .. mean(factor_n)
- полученный предиктор используйте как новый фактор, приджойнив к каждой строке пользователя, и используйте его в машинном обучении

питон-код с использованием pandas

Какая получается сила у нового фактора? Помогает ли он как-либо улучшить, точность, полноту или f1-меру?

1. Выберите "лучший" из имеющихся у Вас классификаторов (по рассчитанным метрикам), с его помощью спрогонзируйте is_fraud_user для файла test.
2. В качестве решения отправьте ipynb-файл с кодом и отдельным файлом столбец нулей и единиц для флага is_fraud_user для файла test.

Хорошо, если у вас получится достичь результата precision=0.75 и/или f1=0.8 на test

Мы постараемся организовать автоматизированный прием решений через Яндекс.Контест. Если не получится, то присылайте столбцы is_fraud_user для test для трех ваших лучших классификаторов в anytask.

-