Доп. задачи анализа данных

Задание 3

Правила:

- Дедлайн **16 мая 23:59**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru` , указав тему письма `"[ad] Фамилия Имя - задание 3"` . Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `3.N.ipynb` и `3.N.pdf` , где `N` - ваш номер из таблицы с оценками.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Для выполнения задания используйте этот ноутбук в качестве основы, ничего не удаляя из него.
- Никакой код из данного задания при проверке запускаться не будет.
- В каждой задаче не забывайте делать **пояснения и выводы**.
- За задание можно получить **10 баллов**.

Одна из важных задач банковских организаций — распознавание мошеннических операции с кредитными картами, например, чтобы клиенты не платили за те операции, которые они не совершают.

Данные
Скачайте данные [https://www.kaggle.com/mlg-ulb/creditcardfraud (https://www.kaggle.com/mlg-ulb/creditcardfraud)](https://www.kaggle.com/mlg-ulb/creditcardfraud)

Набор данных содержит информацию о транзакциях, совершенных европейцами в сентябре 2013 года за два дня. Среди 284 807 транзакций 492 являются мошенническими. Как видно, данные сильно несбалансированы, поскольку неправомерные операции составляют всего 0.172% всех транзакций.

Признаки
Подчекнем важную особенность данных — колонки `V1, ..., V28` содержат только вещественные значения, которые являются результатом PCA трансформации над реальными данными. Реальные данные не предоставлены для выполнения условия конфидицальности клиентов банков. Есть еще 2

признака, которые не были подвергнуты трансформации: `Time` и `Amount`. `Time` — время в секундах между каждой транзакцией и первой транзакцией в датасете. `Amount` — количество денег, участвующих в транзакции.

Таргет
`Class` — принимает значение 1 в случае мошенничества, 0 — иначе.

Ситуация 1: анализ полученных данных

Предположим, нам не известно о том, какие операции являются мошенническими.
Примените изученные методы детектирования *выбросов* для поиска мошеннических операций.
Подберите гиперпараметры методов на основе визуального представения результатов работы метода.
Парметр `contamination` можете ограничить значениями 0.001 и 0.005. Тщательного подбора гиперпарметров от вас не требуется :).

Не забудьте произвести нормализацию признаков. Используйте `RobustScaler из sklearn`, поскольку, возможно не все операции, которые отмечены как не мошеннические, действительно легальные. Поясните почему нормализация необходима.

In [ ]:

```
1
```

Для оценки качества каждого метода посчитайте значения нескольких метрик качества классификации и запишите значения в таблицу. Поясните, почему вы выбрали эти метрики.

In [ ]:

```
1
```

Сделайте выводы.

In [ ]:

```
1
```

Ситуация 2: построение модели распознавания новых мошеннических схем

Теперь рассмотрим методы детектирования новизны.
Разделите выборку на две части следующим образом:

- в первой части данных 50000 легальных транзакций, мошеннических нет, это будет трейн;
- во второй части данных все остальные данные, это будет тест.

Внимание. Нельзя использовать мошеннические операции при обучении или выборе модели. Этих данных нет при построении модели, они появятся потом. Когда они появятся, ваша модель должна быть уже готова их задетектировать.

Нормализуйте данные по трейну. Используйте `RobustScaler из sklearn`, поскольку, возможно не все операции, которые отмечены как не мошеннические, действительно легальные.

In [ ]:

```
1
```

Обучите на трейне `OneClassSVM` .

In [ ]:

```
1
```

Получите результат на тесте и посчитайте метрики качества классификации модели.

In [ ]:

```
1
```

Обучите на трейне автоэнкодер. О том, как написать автоэнкодер, разобрано на первом семинаре по нейронным сетям. Можете попробовать поменять параметры сети, в том числе изменить ширину и количество слоев, заменить функции активации.

In [ ]:

```
1
```

Выберите порог отклонения предсказания от реального значения, по которому можно разграничить легальные и мошеннические транзакции.

In [ ]:

```
1
```

Посчитайте метрики качества классификации в зависимости от порога.

In [ ]:

```
1
```

Сравните результаты. Сделайте выводы.

```
1
```