Гадаем на картах

Можно ли узнать возраст клиента на основе информации о его расходах по карте? Мы подготовили задачу на базе реальных банковских транзакций. Совершенствуя свои продукты, банк использует информацию о пользователях, в том числе и возраст. Это помогает сделать персонализированные продукты, которые удовлетворяют реальные потребности клиентов. Но всегда ли календарный возраст соответствует образу жизни (и покупок) человека?

Точные предсказания

Ваша задача — по информации о расходах клиента банка предсказать, в какую из возрастных групп он попадает. Даны обучающие train данные для построения признаков и обучения моделей, и тестовыеtest данные для проверки алгоритмов. Это специальным образом подготовленная и анонимизированная информация, на которой можно обучать модели, сохраняя полную безопасность реальных данных клиентов. Решением задачи являются предсказания алгоритмов на тестовых данных. Чем меньше ошибок совершит ваш алгоритм, тем выше вы поднимитесь в рейтинге!

### **Данные**

Для решения задачи участникам предоставляется информация о транзакциях клиентов банка. Объемом около 27 000 000 миллионов записей.

Каждая запись описывает одну банковскую транзакцию. Для каждого из ≈20 000 тестовых id, участникам необходимо с помощью обученной модели предсказать — в какую из возрастных групп попадает Клиент.

Мы подготовили два набора данных:

1. Обучающий transactions\_train.csv, в котором для каждой транзакции известна дата, сумма, тип и id клиента;
2. Тестовый transactions\_test.csv, содержащий те же поля:
   * *сlient\_id* – уникальный номер клиента;
   * *trans\_date* – дата транзакции (представляет из себя просто номер дня в хронологическом порядке, начиная от заданной даты);
   * *small\_group* – группа транзакций, характеризующих тип транзакции (например, продуктовые магазины, одежда, заправки, детские товары и т.п.);
   * *amount\_rur* – сумма транзакции (для анонимизации данные суммы были трансформированы без потери структуры).

На базе данных файлов можно строить различные признаки, которые характеризуют возрастные группы.

Целевая переменная для обучающего датасета находится в файле train\_target.csv. В нем содержится информация о Клиенте и метка возрастной группы, к которой он относится:

* *client\_id* – уникальный номер Клиента (соответствует *client\_id* из файлаtransactions\_train.csv);
* *bins* – метка возраста. В файлe test.csv вам надо предсказать для указанных *client\_id* соответствующую метку группы возраста.

Участникам также предоставлен информационный файл small\_group\_description.csv, который содержит расшифровку типов транзакций.

### **Формат решений**

Для каждого примера из тестового набора необходимо предсказать возрастную группу к которой относится клиент. В систему необходимо предоставить для проверки CSV-файл с предсказаниями, он должен содержать две колонки:

1. *client\_id* — идентификатор клиента;
2. *bins* — возрастная группа.

Пример выходных данных:

  client\_id,bins  
0,0  
7,1  
9,0  
10,2  
11,1  
15,3  
...

Задача представляет из себя мультиклассовую классификацию (4 класса – от 0 до 3). Качество решения считается как доля верно угаданных меток возраста по всем тестовым примерам - accuracy.

Для решения удобнее всего использовать язык программирования Python, так как для него есть большое число библиотек для анализа данных: NumPy, Pandas, SciKit-Learn и другие. В качестве инструмента разработки — интерактивную среду Jupyter.

Финал

Ваше решение вошло в число лучших по рейтингу? Тогда вас ждет финал! 100 лучших участников приедут в начале марта в Москву. Здесь их ждут 3 дня кодинга в командах из 2 человек и защита проекта перед практикующими data scientist'ами. Участники команд, предложивших лучшие решения, получат сертификаты на образование номиналом 100 000 рублей, новую технику и предложения о стажировке от Сбербанка!