**Техническое задание по проекту: Построение ML-продукта для выявления и оптимизации платежей преподавателей сервиса Repetit.ru**

**Содержание**

[Заказчик](#__RefHeading___1)

[Описание проекта](#__RefHeading___2)

[Сервис передает контакты клиента (ученика) репетитору. Если репетитор начинает заниматься с учеником, то он должен платить сервису комиссию от каждого занятия. Но в реальности так происходит не всегда. Иногда, это из-за того, что репетитор звонит по телефону и ему просто не отвечают. Некоторые репетиторы плохо договариваются о занятиях или обманывают. Сервис теряет деньги каждый раз, когда отдаёт заявку неэффективному репетитору. Заказчику нужно как можно раньше понять, что репетитор недобросовестный или мошенник, чтобы отключить его от сервиса и отдавать заявки ответственным репетиторам.](#__RefHeading___3)

[Признаки обмана, выявленные при ручной проверке](#__RefHeading___4)

[Описание данных](#__RefHeading___5)

[Используемый стек технологий](#__RefHeading___6)

[Срок реализации](#__RefHeading___7)

[План реализации:](#__RefHeading___8)

[Ожидаемый результат](#__RefHeading___9)

[План вебинаров Мастерской](#__RefHeading___10)

[Полезные материалы](#__RefHeading___11)

## **Заказчик**

Сервис по подбору репетиторов Repetit.ru

## **Описание проекта**

## Сервис передает контакты клиента (ученика) репетитору. Если репетитор начинает заниматься с учеником, то он должен платить сервису комиссию от каждого занятия. Но в реальности так происходит не всегда. Иногда, это из-за того, что репетитор звонит по телефону и ему просто не отвечают. Некоторые репетиторы плохо договариваются о занятиях или обманывают. Сервис теряет деньги каждый раз, когда отдаёт заявку неэффективному репетитору. Заказчику нужно как можно раньше понять, что репетитор недобросовестный или мошенник, чтобы отключить его от сервиса и отдавать заявки ответственным репетиторам.

Сейчас эта задача решается ручным просмотром сотрудниками или никак.

**Задачи:**

Разработать модель, которая по имеющейся информации о репетиторе и проведенных им занятий будет классифицировать репетиторов на платящих и неэффективных. Оценка качества модели будет производиться с использованием метрики F1.

### 

### **Признаки обмана, выявленные при ручной проверке**

Заказчик произвел ручную проверку репетиторов на предмет мошенничества (звонки клиентам и так далее). Выяснилось, что 20% обманывают нас так или иначе.

* Несоответствие цены в заявке, в анкете репетитора и реальной
* 1 занятие на ученика (соотношение к общему кол-ву занятий)
* Разная стоимость с разными учениками
* Разная стоимость уроков с одним учеником
* Стоят занятия в расписании, а оплат нет
* Когда у репетитора по разным ученикам оплаты рядом по времени, то он сам оплачивает
* Как часто заходит в приложение - если редко, то сам оплачивает
* Статус заявки договорились о занятиях, оплат нет более недели (и не перенесено)
  + status в orders = 9
  + lessons с amount\_payed > 0 за последнюю неделю
* Ученик не отмечен как завершенный, а оплат нет (пример SQL запроса ниже)

SELECT \* [lesson\_course].id , (select count(lessons.id) from lessons where lessons.lesson\_course\_id=lesson\_course.id and lessons.amount\_paid is not NULL and lessons.lesson\_date>getdate()-30) as 'оплат за последний месяц' FROM [main].[dbo].[lesson\_course] join orders on orders.id=lesson\_course.order\_id where orders.flags = 8 and orders.status\_id in (6,14) and lesson\_course.is\_active=1 and (lesson\_course.suspend\_till\_date is null or lesson\_course.suspend\_till\_date <getdate())

* Репетитор отчитался, что провёл платное занятие, оплаты нет (пример SQL запроса ниже)

lesson\_course.id , (select count(lessons.id) from lessons where lessons.lesson\_course\_id=lesson\_course.id and lessons.amount\_paid is not NULL and lessons.lesson\_date>getdate()-30) as 'оплат за последний месяц' FROM [main].[dbo].[lesson\_course] join orders on orders.id=lesson\_course.order\_id where orders.flags = 8 and lesson\_course.is\_active=1 and (lesson\_course.suspend\_till\_date is null or lesson\_course.suspend\_till\_date <getdate()) and (select count(lessons.id) from lessons where lessons.lesson\_course\_id=lesson\_course.id and lessons.amount\_paid is not NULL and lessons.lesson\_date>getdate()-30)=0 and (select count(id) from reports where reports.order\_id=orders.id and [description] = 'Проведено первое занятие' and [comments] not like '%Стоимость первого занятия - 0%' )>0 order by orders.id desc

* Цена ниже 500 р. в регионах, ниже 700 в мск

## **Описание данных**

**Информация о репетиторах (teacher\_info.feather)**

* id - айди репетитора
* reg\_date - дата регистрации
* birth\_date - дата рождения
* teaching\_start\_date - дата начала первого занятия
* is\_email\_confirmed - подтвержден ли e-mail адресс
* lesson\_duration - продолжит урока
* lesson\_cost - стоимость урока
* is\_display - показывается в каталоге 'что именно значит'
* last\_visited - послеждний визит?
* is\_pupils\_needed - открыт для заявок
* is\_cell\_phone\_confirmed - подтвержден ли номер телефона
* area\_id - регион
* sex - пол? почему 3
* orders\_allowed - разрешено назначать на заявки
* review\_num - отзывы

**Статистика по репетиторам и таргет (teachers.feather)**

* id - айди репетитора
* lessons\_delivered - поставлено уроков
* mean\_lesson\_price - средняя стоимость уроков
* lessons\_given - оплачено уроков
* lessons\_started\_fraction - процент начала занятий
* lessons\_per\_pupil - занятий на ученика
* money\_recieved - получено денег
* blocked - целевой признак (active/blocked)

**Ученики (lesson\_course.feather)**

* Id - айди
* client\_id - айди ученика
* teacher\_id - айди репетитора
* order\_id - айди заявки
* lesson\_place - занятия онлайн или офлайн
* lesson\_price - цена
* is\_active - идут ли занятия, на паузе, завершены
* lesson\_duration - продолжительность урока
* date\_updated
* suspend\_till\_date

**Занятия (lessons.feather)**

* Id - айди
* lesson\_course\_id - айди ученика
* lesson\_date - дата
* time\_from - время от
* time\_to - время до
* home\_task - дз
* is\_regular - автоматически повторяющееся занятие
* amount\_to\_pay - стоимость
* amount\_paid - оплачено

**Цены на занятия репетиторов (teacher\_prices.feather)**

* date\_update - дата обновления цен
* teacher\_id - айди репетитора
* subject\_id - айди предмета
* price - цена занятий у себя
* price\_external - цена занятий на выезде
* price\_remote - цена онлайн занятий

**Заявки (orders.feather)**

* order\_date - дата создания
* subject\_id - предмет
* purpose - цель занятий
* lesson\_price - цена
* lesson\_duration - желаемая проодолжительность урока
* home\_metro\_id - ближайшее метро
* add\_info - доп инфо
* start\_date
* working\_teacher\_id
* status\_id - оплачена ли заявка (значения 6 и 13 говорят о факте оплаты заявки)
* comments
* amount\_to\_pay
* planned\_lesson\_number - клиент планирует N занятий
* first\_lesson\_date - дата 1 занятия
* creator\_id - кто создал заявку (id сотрудника или клиента)
* pupil\_category\_new\_id - возраст ученика
* lessons\_per\_week - занятий а неделю
* minimal\_price
* teacher\_sex - пол репетитора
* teacher\_experience\_from - опыт репетитора от
* teacher\_experience\_to- опыт репетитора до
* lesson\_place\_new - онлайн, у ученика, у учителя
* pupil\_knowledgelvl -уровень знаний ученика
* teacher\_age\_from - желаемый возраст репеитора от
* teacher\_age\_to - желаемый возраст репеитора от
* chosen\_teachers\_only - не предлагать репетиторов кроме выбранных самостоятельно
* no\_teachers\_available - на заявку нет подходящих репов
* source\_id - где создана заявка (какая часть сайта, не регион)
* original\_order\_id - дублем какой заявки является эта заявка
* client\_id - айди клиента
* additional\_status\_id
* max\_metro\_distance - максимально готов ехать от метро
* estimated\_fee
* payment\_date
* test\_group - аб тесты
* is\_display\_to\_teachers - хочет ли клиент получать отклики репетиторов

## **Используемый стек технологий**

* Python
* Pandas
* Numpy
* Matplotlib
* Scikit-learn
* LightGBM
* CatBoost

## **Срок реализации**

Срок реализации проекта - 3 недели с момента старта Мастерской

## **План реализации**:

* загрузка и ознакомление с данными,
* предварительная обработка и отбор полезных признаков,
* полноценный разведочный анализ,
* разработка новых синтетических признаков,
* отбор финального набора обучающих признаков.
* выбор и обучение моделей,
* итоговая оценка качества предсказания лучшей модели,
* анализ важности ее признаков.
* подготовка отчета по исследованию

## **Ожидаемый результат**

Тетрадь с решением задачи (описание проекта, исследование, методы решения)

## **План вебинаров Мастерской**

1. Вводная встреча
2. QA встреча
3. QA встреча
4. Финальная встреча

## **Полезные материалы**

CatBoost - https://habr.com/ru/companies/otus/articles/778714/