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

Содержание

Заказчик
Описание проекта
Описание данных
Используемый стек технологий
Срок реализации
План реализации
Ожидаемый результат
План вебинаров Мастерской
Полезные материалы

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
- last_visited - послежднний визит
- is_pupils_needed - открыт для заявок
- is_cell_phone_confirmed - подтвержден ли номер телефона
- area_id - регион
- sex - пол
- 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/