

1. Согласно принципу SMART, попробуйте расписать кейс кредитного скоринга, разбиравшийся на одном из предыдущих курсов.

S – specific.

Разработать бинарный классификатор для предсказания дефолта клиента по его признаковому описанию.

M – measurable.

Определяющая метрика F1 позитивного класса бинарного целевого признака (дефолт клиента) должна быть больше 0.5 при точности и полноте позитивного класса также более 0.5.

A – actionable.

1. Загрузка, обработка данных и автоматизация этих процессов:
 - 1.1. Визуальный и статистический анализ признаков в зависимости от целевого признака
 - 1.2. Определение и задание типов признаков (бинарный, категориальный, вещественный)
 - 1.3. Разбиение данных на обучающую, тестовую и/или валидационную выборки
 - 1.4. Обработка пропущенных значений (самое частое, среднее или нехарактерное значение и т.д.)
 - 1.5. Устранение выбросов и спорных, нехарактерных объектов
 - 1.6. Синтез новых признаков
 - 1.7. Отбор значимых признаков
 - 1.8. Балансировка классов в целевом признаке в обучающей выборке (oversampling, undersampling, reweighting или не требуется)
 - 1.9. Масштабирование данных (нормализация, стандартизация или не требуется)
2. Построение модели и проверка ее на тестовой и/или валидационной выборках:
 - 2.1. Построение и проверка базовой модели
 - 2.2. Поиск наилучшего по заданной метрике классифицирующего алгоритма (модели, гиперпараметров)
 - 2.3. Поиск наилучшей по заданной метрике комбинации способов обработки данных из п.1
3. Проверка работы модели «в боевом режиме» — в том числе и сторонними метриками. Анализ количества и динамики ошибок первого и второго рода.

R – relevant.

Разрабатываемый классификатор является важным и полезным инструментом для поддержки принятия решения при выдаче клиенту кредита. Кроме того, он может быть ключевой частью других сервисов.

T – time bound.

Выделена неделя после последнего вебинара, чтобы предоставить решение в виде Jupiter-ноутбука.

2. Предложить перечень индикаторов оттока для клиента каршеринга. Какие данные мы бы взяли в датасет, предполагая, что в нашем распоряжении есть какие угодно сведения?

- Дата регистрации пользователя
- Дата первого входа в приложение и на сайт сервиса
- Дата рождения пользователя
- Населенный пункт, регион места рождения и текущего проживания пользователя
- Место работы или учебы и должность
- Вся ли информация о пользователе заполнена (сколько и какие поля не заполнены)
- Текущие дата и время (индикаторы: сезон, выходные, праздники)
- Индикаторы, когда пользователь в среднем пользуется сервисом (время дня, будни, выходные, праздники, сезон)
- Дата получения водительских прав пользователем (срок)
- Водительский стаж (сколько по факту водит пользователь)
- Количество арендованных машин к текущему моменту
- Динамика пользования сервисом (количество аренд, входов в приложение или на сайт и т.д. за последний указываемый период)
- Есть ли у пользователя автомобиль (их количество; свой и/или служебный)
- Тарифный план (поминутный или долгосрочный абонемент, смены тарифа)
- Дата последней аренды
- Дата последнего входа в приложение и на сайт сервиса
- Дата последней транзакции
- Общая и средняя сумма транзакций
- Тип и класс арендуемого автомобиля (премиум, грузовой и т.п.)
- Тип прав пользователя
- Общая и средняя длительность поездок
- Регион эксплуатации арендуемых автомобилей (предпочитаемые пользователем районы и маршруты, есть ли отклонения от них и т.д.)
- Аварии, аварийные ситуации, нарушения скоростного режима и ПДД, поломки автомобиля за все время (их количества) и во время последней аренды (был ли факт)
- Жалобы на состояние автомобиля (от пользователя и «после» пользователя)
- Отзывы на сервис от пользователя (положительные, отрицательные; на то он обычно жалуется)
- Участие в скидках, акциях и программах лояльности
- Участие в конкурирующих каршеринговых сервисах (желательно те же данные — т.е. насколько пользователь предпочитает именно данный сервис).