ПРОГНОЗИРОВАНИЕ ОТТОКА ПОЛЬЗОВАТЕЛЕЙ

ДАТАСЕТ

МФТИ

- Данные разного типа
  - числовые
  - номинальные
  - порядковые
- Временные ряды

НЕСБАЛАНСИРОВАННАЯ ВЫБОРКА

МФТИ

- Доля целевого класса может быть намного меньше доли нецелевого класса (0.1% vs 99.9%)
- Несбалансированность выборки может негативно сказаться на качестве модели
- Важно заметить это в процессе построения модели!

REWEIGHTING

МФТИ

- Задать веса для объектов таким образом, чтобы:
  - Скомпенсировать количество объектов меньшего класса их важностью
  - Задать стоимость ошибки классификации разного рода

OVERSAMPLING

МФТИ

- Сгенерировать больше объектов меньшего класса:
  - Дублирование объектов
  - Генерация новых объектов путем изменения некоторых признаков существующих объектов
  - Генерация новых объектов на основе нескольких существующих объектов

UNDERSAMPLING

МФТИ

- Исключить из обучения объекты преобладающего класса:
  - Удаление из выборки случайных объектов преобладающего класса
  - Удаление из выборки групп схожих объектов из преобладающего класса

ОБУЧЕНИЕ МОДЕЛИ

МФТИ

- Обучение на данных, доступных НЕ только за исторический период
- Контроль обучения на данных из будущего
- Контроль переобучения

КРОСС-ВАЛИДАЦИЯ

МФТИ

- По объектам
- По времени

ПОДБОР ПАРАМЕТРОВ

МФТИ

- Используем кросс-валидацию
- Сразу фиксируем hold-out dataset
- Их может быть несколько
- Используем для финальной проверки решения

ОЦЕНКА КАЧЕСТВА

МФТИ

- Одна целевая метрика
- Оффлайн метрика совпадает или коррелирует с целевой метрикой
- Хорошо оценивать модель «скользящим окном» по времени

ПЕРСПЕКТИВЫ УЛУЧШЕНИЯ МОДЕЛИ

МФТИ

- На каких группах объектов модель ошибается?
- Является ли инвестиция в дальнейшее улучшение модели экономически оправданной?

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

МФТИ

- Какие факторы внесли наибольший вклад в модель?
- Гипотезы относительно причин оттока пользователей?
- Какие объекты классифицируются наиболее/наименее уверенно?
- Какие еще данные могли бы быть полезны?

РЕГУЛЯРНОЕ ОБНОВЛЕНИЕ МОДЕЛИ

МФТИ

- Как меняется качество модели во времени?
- Как быстро она «протухает»?
- Сколько времени занимает переобучениие модели?
- Сколько времени требуется на переключение с одной модели на другую?

МОНИТОРИНГ КАЧЕСТВА

МФТИ

- Изменились ли данные?
- Изменилось ли качество модели?
- Хорошо оценивать модель с разных сторон с помощью набора метрик

ВАЖНО

МФТИ

- Оценивать качество решения задачи на всех этапах
- Заранее продумать список потенциальных «узких» мест