DATA

УЛУЧШЕНИЕ КАЧЕСТВА МОДЕЛЕЙ

ПЛАН ВЕБИНАРА

- Повторим основные понятия, изученные в рамках темы
- Поделаем упражнения на понимание терминов
- Порешаем задачку из собеседования
- Узнаем, как кодировать категориальные переменные через значения целевого признака
- Разберем проблемы, возникающие, при выполнении задачи
- Разберем вопросы по теме

МЕТОДЫ БОРЬБЫ С ДИСБАЛАНСОМ КЛАССОВ

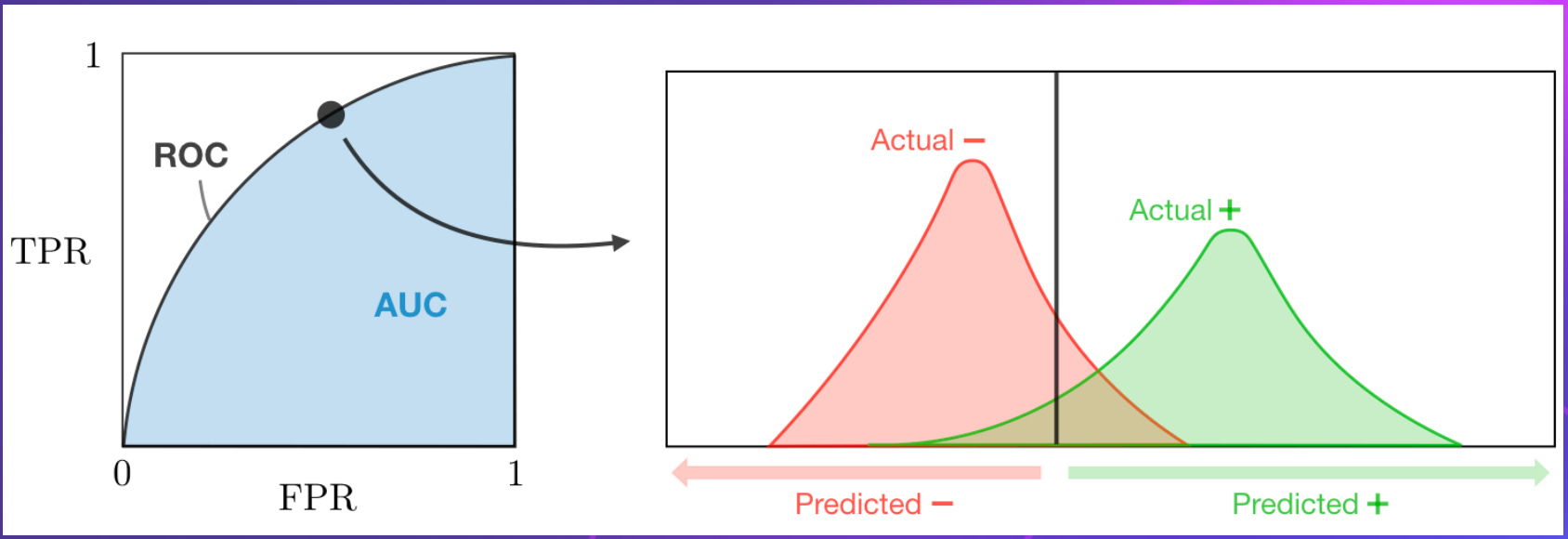
- Как поменяется точность и полнота у модели, решающей задачу бинарной классификации, если увеличить порог классификации?
- Каким будет вес класса 1 в модели, решающей задачу бинарной классификации, если в обучающей выборке 945 объектов класса 0 и 315 объектов класса 1?
- Что лучше применить upsampling или downsampling, если в обучающей выборке 456 объектов класса 0 и 512 объектов класса 1?

TARGET ENCODING

- Как рассчитывается
  - Для задачи регрессии используется среднее значение целевого признака по данному значению категориального признака
  - Для задачи бинарной классификации - вероятность 1го класса для данного значения категориального признака
- Минус - утечка целевого признака в переменную, как бороться?
  - Leave-One-Out - при расчете показателя для конкретного наблюдения, его значение не используется
  - James-Stein-Encoder - рассчитывается средневзвешенное между значением для данного категориального признака и значением по всей выборке
- Целесообразно использовать, если много значений у категориального признака

ROC-AUC И ЗАДАЧА ИЗ СОБЕСЕДОВАНИЯ

Одна из интерпретаций: вероятность того, что случайно выбранный объект класса 1 будет отранжирован моделью выше, чем случайно выбранный объект класса 0



? Как улучшить алгоритм, если ROCAUC = 0.1

ПРОБЛЕМЫ, ВОЗНИКАЮЩИЕ ПРИ ВЫПОЛНЕНИИ ЗАДАЧИ

- Подтверждается, что в классах сильный дисбаланс, но ни один из изученных методов не применяется. Объектов 0-го класса в 4 раза больше, чем объектов 1го. Самый простой вариант - взвешивать наблюдения в лоссе class_weight=‘balanced'
- При ОНЕ-кодировании не выбрасывается один из итоговых столбцов. Необходимо использовать либо pd.get_dummies(drop_first=True), либо sklearn.preprocessing.OneHotEncoder(drop=‘first’)
- Стандартизацию необходимо настраивать только на обучающей выборке. Преобразование тестовой части нужно проводить уже настроенным скейлером.
- Нельзя использовать LabelEncoder и OrdinalEncoder с линейными моделями.

ВОПРОСЫ ПО ТЕМЕ И ДОМАШНЕМУ ЗАДАНИЮ