Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Программная инженерия. Разработка ПО (Python для продвинутых специалистов. Машинное обучение)

Модуль: Предобработка данных и машинное обучение

Лекция 7: Кодирование категориальных данных

Дата: 05.06.2025

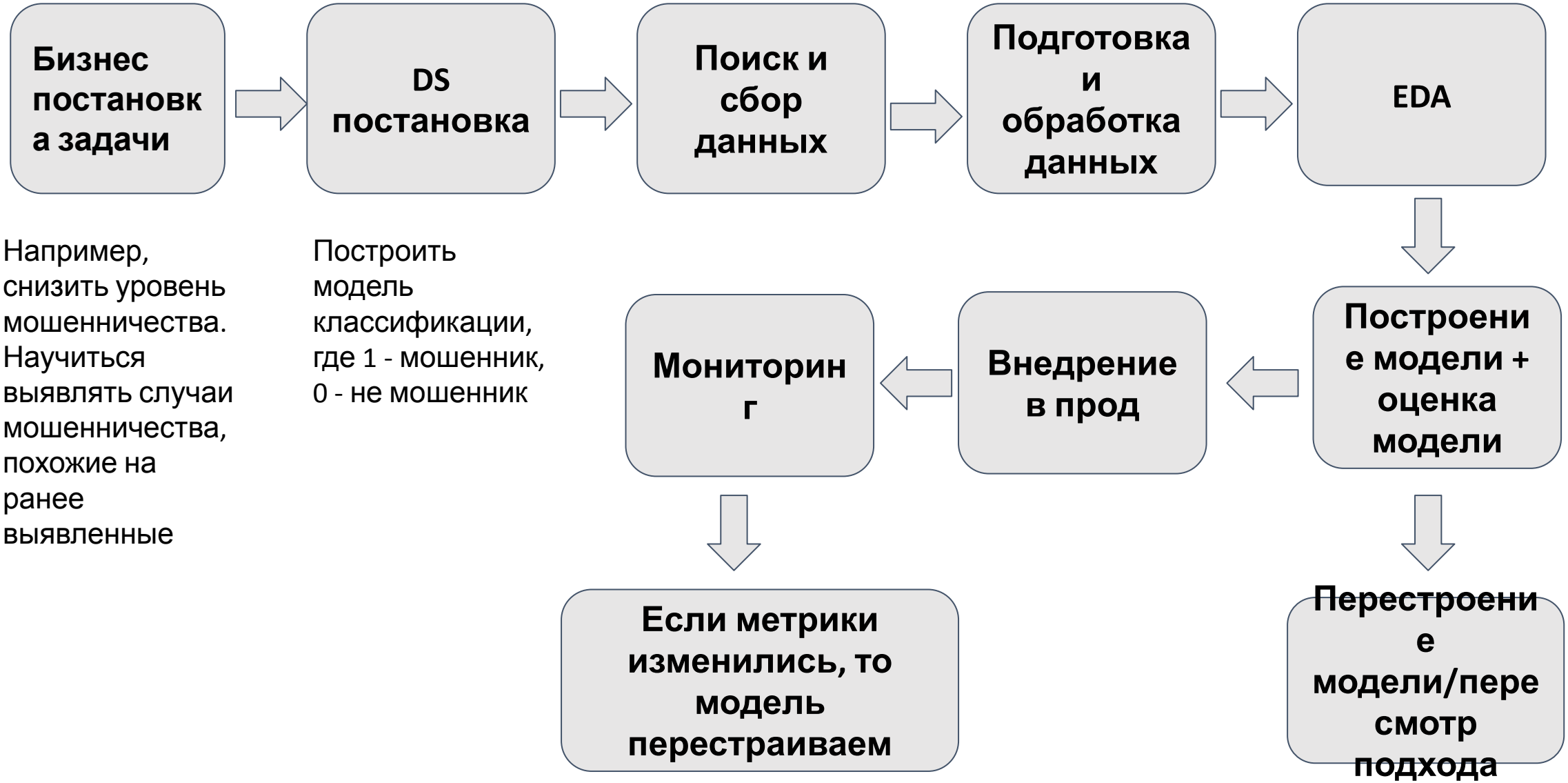
Содержание лекции

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

- Процесс разработки модели ML
- Обработка категориальных данных
- Практика

Процесс разработки модели ML

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

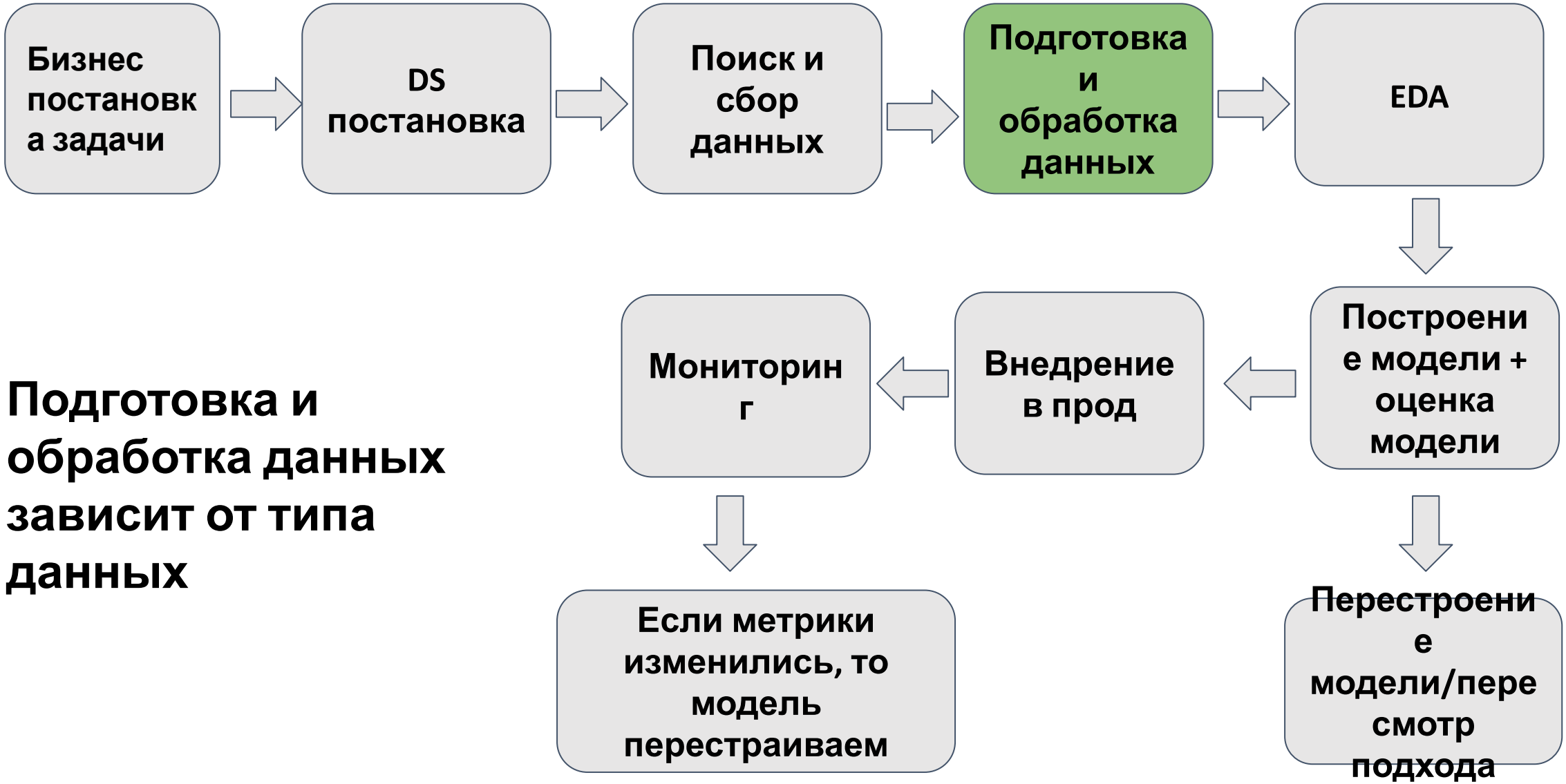


Например, снизить уровень мошенничества. Научиться выявлять случаи мошенничества, похожие на ранее выявленные

Построить модель классификации, где 1 - мошенник, 0 - не мошенник

Процесс разработки модели ML

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС



Подготовка и обработка данных зависит от типа данных

Типы переменных(features)

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Числовые (Quantitative / Numeric):

- **Непрерывные (Continuous)** — могут принимать любое значение из диапазона.
  - Примеры: рост, вес, температура, цена.
- **Дискретные (Discrete)** — принимают только целочисленные значения.
  - Примеры: количество детей, число покупок, уровень образования.

2. Категориальные (Categorical):

- **Номинальные (Nominal)** — категории без упорядочивания.
  - Примеры: пол (м/ж), цвет (красный, синий), страна.
- **Порядковые (Ordinal)** — категории с естественным порядком.
  - Примеры: уровень образования (начальный < средний < высший), оценка по шкале (низкий, средний, высокий).

Какие могут быть проблемы у числовых переменных?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

- Пропущенные значения
- Выбросы/аномальные значения
- Разные шкалы, то есть в выборке есть

  один признак (feature) в (1, 5),

  а другой в (100, 5_000)

ID магазина	площадь	количество этажей	в ТЦ?	доход от магазина
1	1000	1	1	1000000
2	1569	2	0	200000
3	870	1	0	300000
4	2000	2	0	500000
5	900	1	1	600000
6	850	1	1	1000000
7	1700	2	1	200000
8		2	1	300000
9		2	0	500000
10	700	1	0	600000

Какие могут быть проблемы у категориальных переменных?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

- Пропущенные значения
- Выбросы/аномальные значения
- Сами категориальные значения - это огромная проблема

!!!Компьютер не умеет умножать/складывать число на текстовое значение

ID магазина	месторасположение в городе	площадь	количество этажей	в ТЦ?	доход от магазина
1	в центре	1000	1	1	1000000
2	на окраине	1569	2	0	200000
3	за пределами города	870	1	0	300000
4	в 10 км от центра города	2000	2	0	500000
5	за пределами города	900	1	1	600000
6	в 10 км от центра города	850	1	1	1000000
7	на окраине	1700	2	1	200000
8	в центре	1400	2	1	300000
9	в центре	1300	2	0	500000
10	в 10 км от центра города	700	1	0	600000

Какие могут быть проблемы у категориальных переменных?

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Самые популярные методы:

- one-hot-encoding (добавляем признаки бинарные признаки)
- mean target encoding (заменим каждую категорию на среднее значение целевой переменной по всем объектам этой категории)
- frequency encoding (заменим каждую категорию частотой появления)

ID магазина	месторасположение в городе	площадь	количество этажей	в ТЦ?	доход от магазина
1	в центре	1000	1	1	1000000
2	на окраине	1569	2	0	200000
3	за пределами города	870	1	0	300000
4	в 10 км от центра города	2000	2	0	500000
5	за пределами города	900	1	1	600000
6	в 10 км от центра города	850	1	1	1000000
7	на окраине	1700	2	1	200000
8	в центре	1400	2	1	300000
9	в центре	1300	2	0	500000
10	в 10 км от центра города	700	1	0	600000

Предобработка данных. Категориальные признаки

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

one-hot-encoding

Сначала определяем сколько групп и что за группы в данных (всего 4 значения): в 10 км от центра города, в центре, за пределами города, на окраине

	bin_в 10 км от центра города	bin_в центре	bin_за пределами города	на окраине
в 10 км от центра города	1	0	0	0
в центре	0	1	0	0
за пределами города	0	0	1	0
на окраине	0	0	0	1

Предобработка данных. Категориальные признаки

one-hot-encoding

один признак можно точно вычислить по другим трем, например,

bin на окраине = 1 - bin_в 10 км от центра города - bin_в центре - bin_за пределами города

	bin_в 10 км от центра города	bin_в центре	bin_за пределами города	bin_на окраине
в 10 км от центра города	1	0	0	0
в центре	0	1	0	0
за пределами города	0	0	1	0
на окраине	0	0	0	1

Предобработка данных. Категориальные признаки

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

one-hot-encoding

Преобразует каждую категорию в отдельный бинарный признак (0/1)

Подходит для: категориальных признаков с небольшим числом уникальных значений

ID магазина	месторасположение в городе	bin_в 10 км от центра города	bin_в центре	bin_за пределами города	площадь	количество этажей	в ТЦ?	доход от магазина
1	в центре	0	1	0	1000	1	1	1000000
2	на окраине	0	0	0	1569	2	0	200000
3	за пределами города	0	0	1	870	1	0	300000
4	в 10 км от центра города	1	0	0	2000	2	0	500000
5	за пределами города	0	0	1	900	1	1	600000
6	в 10 км от центра города	1	0	0	850	1	1	1000000
7	на окраине	0	0	0	1700	2	1	200000
8	в центре	0	1	0	1400	2	1	300000
9	в центре	0	1	0	1300	2	0	500000
10	в 10 км от центра города	1	0	0	700	1	0	600000

Предобработка данных. Категориальные признаки

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

one-hot-encoding

И в итоге вместо 1 переменной получаем N-1 переменную, где N - количество значений переменной

Что будет, если у переменной очень много значений, например 100? мы создадим вместо 1 переменной 99?

А если в наших данных таких категориальных переменных несколько, например 3?

мы создадим 99 + 99 + 99 = 297 дополнительных переменных

Что же делать?

Не использовать категориальные переменные? :((((

Предобработка данных. Категориальные признаки

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

mean target encoding (заменим каждую категорию на среднее значение целевой переменной по всем объектам этой категории)

месторасположение в городе	AVERAGE of доход от магазина
в 10 км от центра города	700000
в центре	600000
за пределами города	450000
на окраине	200000

месторасположение в городе	mean_месторасположение в городе	площадь	количество этажей	в ТЦ?	доход от магазина
в центре	600000	1000	1	1	1000000
на окраине	200000	1569	2	0	200000
за пределами города	450000	870	1	0	300000
в 10 км от центра города	700000	2000	2	0	500000
за пределами города	450000	900	1	1	600000
в 10 км от центра города	700000	850	1	1	1000000
на окраине	200000	1700	2	1	200000
в центре	600000	1400	2	1	300000
в центре	600000	1300	2	0	500000
в 10 км от центра города	700000	700	1	0	600000

Предобработка данных. Категориальные признаки

Но и здесь нас ждут неожиданности :)

при обучении моделей mean target encoding необходимо рассчитывать на отложенной выборке, не на всей

то есть обучать на трейне, а к тесту применять

месторасположение в городе	mean_месторасположение в городе	площадь	количество этажей	в ТЦ?	доход от магазина
в центре	600000	1000	1	1	1000000
на окраине	200000	1569	2	0	200000
за пределами города	450000	870	1	0	300000
в 10 км от центра города	700000	2000	2	0	500000
за пределами города	450000	900	1	1	600000
в 10 км от центра города	700000	850	1	1	1000000
на окраине	200000	1700	2	1	200000
в центре	600000	1400	2	1	300000
в центре	600000	1300	2	0	500000
в 10 км от центра города	700000	700	1	0	600000

Процесс предобработки данных

ПЕРЕДОВАЯ ИНЖЕНЕРНАЯ ШКОЛА УНИВЕРСИТЕТА ИННОПОЛИС

Делим данные на train и test!

На трейне **анализ выбросов.** Выбор стратегии обработки **выбросов**. Обучение алгоритма обработки выброса на трейне, к тесту только применяем полученный результат

На трейне **анализ пропусков.** Выбор стратегии обработки **пропусков**. Обучение алгоритма обработки пропусков на трейне, к тесту только применяем полученный результат

На трейне анализ **категориальных признаков.** Выбор стратегии обработки **кат признаков.** Обучение алгоритма на трейне, к тесту только применяем полученный результат

Приведение к единой шкале**.** Обучение алгоритма на трейне, к тесту только применяем полученный результат

Данные готовы!

Передовые инженерные школы

МИНОБРНАУКИ РОССИИ

УНИВЕРСИТЕТ ИННОПОЛИС

онлайн университет

Спасибо за внимание