ДОПОЛНИТЕЛЬНОЕ ЗАДАНИЕ

Faculty of Computer Science HSE University + Я

КОМАНДА «ТРИ КОТА И ОДНА КОШЕЧКА», 10 КЛАСС

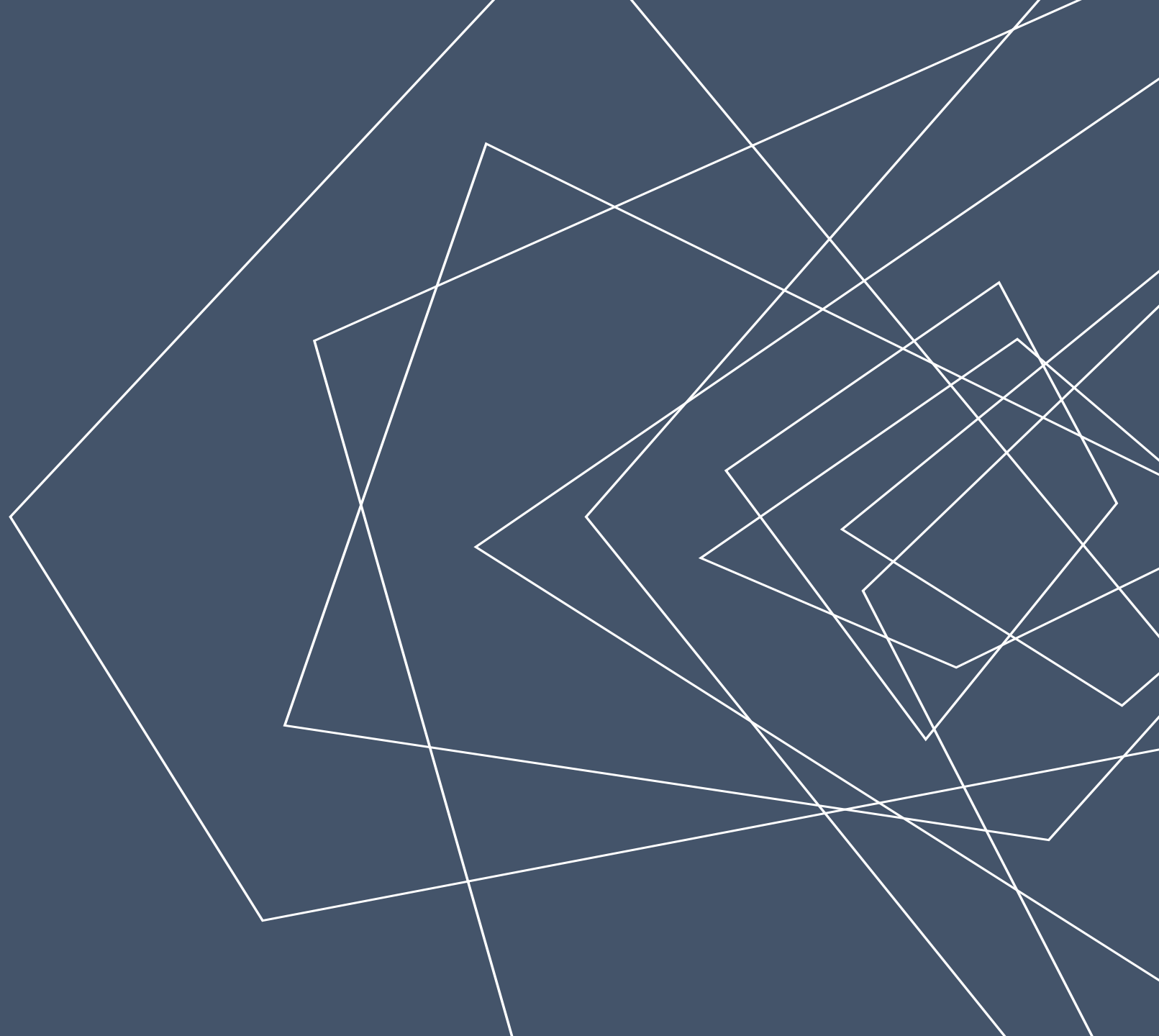
Уханова Екатерина Максимовна
Астрахань
Team lead, market analyst

Локтионова Екатерина Михайловна
Курск
Design, data engineering

Карпович Андрей Евгеньевич
Санкт-Петербург
ML-development

Волкова Екатерина Сергеевна
Ярославль
Data engineering

ЧАСТЬ 1



ОБРАБОТКА ОТСУТСТВУЮЩИХ ЗНАЧЕНИЙ

- ☑ Для обработки отсутствующих значений воспользуемся как удалением, так и заполнением
- ☑ Удалять будем столбцы, слабо коррелирующие с параметром цены, и при этом имеющие большое количество пропусков
- ☑ Те параметры, данные по которым отсутствуют лишь в малом количестве строк (<5%), мы не удаляем, а удаляем сами строки, в которых пропущены данные
- ☑ Остается три столбца: num_room, state и life_sq. Так как в них находится большое количество пропущенных данных, но с ценой они сильно коррелируют, будем заполнять пропуски регрессией для того, чтобы не потерять большую часть датасета.
- ☑ Как итог у нас получился датасет без пустых значений

ВЫЯВЛЕНИЕ АНОМАЛИЙ

- ☑ Для выявления аномалий воспользуемся функциями mean и std, которые показывают нам среднее значение параметров и их стандартное отклонение
- ☑ Будем называть аномальными значения, превосходящие среднее на 3 стандартных отклонения
- ☑ Мы можем видеть, что большинство данных, имеющих аномалии – это информация про район, которая скорее всего аномальна не из-за выбросов, а банально из-за какого-то фактора, такого как «церковь находится очень далеко в этом районе». Поэтому будем перебирать значения в сторону повышения пока мы не найдем threshold, при котором эти факторы пропадут. Такое число это 15.
- ☑ Просто удалим строки с аномальными значениями (так как их не так много)

ОБРАБОТКА ЛИШНИХ ЗНАЧЕНИЙ

- ☑ Лишние значения это те, которые сильно коррелируют между собой. Такие значения могут понизить качество обучения модели
- ☑ Для нахождения сильно коррелирующих значений воспользуемся функцией .corr для нахождения корреляции между всеми данными.
- ☑ Высокая степень корреляции – это 0.5 и выше, поэтому отделим такие пары значений
- ☑ Обработка лишних значений далее на данном этапе безполезна, так как при дальнейшем отборе признаков по важности(см. далее) не будет признаков с высокой корреляцией между собой

ЧАСТЬ 2

СБАЛАНСИРОВАННОСТЬ ДАННЫХ

- ☑ Для измерения сбалансированности будем использовать skewness и kurtosis – параметры показывающие симметричность и распределенность данных по оси.
- ☑ В нашем случае данные сбалансированны если у них низкий kurtosis и skewness близкий к 0
- ☑ Мы можем заметить, что большинство наших данных очень плохо сбалансированы, что может быть связано с выбором объектов недвижимости

БАЗОВЫЙ ОТБОР ПРИЗНАКОВ

- ☑ Для отбора признаков будем использовать встроенную в модель случайного леса функцию важности параметров
- ☑ Сначала обучим модель на всех признаках, затем отберем важные и переобучим модель на новых признаках
- ☑ Возьмем признаки с важностью больше 0.01
- ☑ Точность понизилась незначительно, зато скорость тренировки значительно повысилась

СТАТИСТИКИ

- ☑ Подсчитаем статистику по средней стоимости квадратного метра жилья по годам.
- ☑ Скорее всего, данные уже были пересчитаны под инфляцию, о чем не было сказано нигде в задании, но тенденция цен соблюдается, с небольшим отклонением вниз. Возможно датасет не берет во внимание очень дорогие квартиры с более высокой ценой квадратного метра

	2011	2012	2013	2014	2015
Данные irn.ru, $\frac{Р}{м^2}$	156 789	163 385	168 003	189 504	177 010
Наши данные, с пересчетом под инфляцию, $\frac{Р}{м^2}$	218 346	219 372	219 795	217 920	137 801
Наши данные, без пересчета под инфляцию, $\frac{Р}{м^2}$	123 153	123 132	127 558	154 298	142 174

ВЫВОДЫ

Предоставленный датасет, даже после всех обработок, имеет достаточный объем, чтобы качественно тренировать модель (коэффицент детерминации 0.78)

Вполне возможно, что если проводить исследование без пересчета на инфляцию, которая по всей видимости уже была сделана, данные получились бы точнее, но, к сожалению, у нас не остается времени это сделать

СПАСИБО ЗА ВНИМАНИЕ