

Тема 9. Предварительная обработка данных

Практические задания для самостоятельного выполнения

Задания выполняются по вариантам. Формулировки заданий общие для всех вариантов; конкретный набор данных, подлежащий изучению, выбирается в соответствии с номером своего варианта.

Результаты выполнения заданий необходимо представить в виде двух файлов:

- 1) ноутбук в формате *ipynb*, содержащий программный код, результаты его выполнения, а также все необходимые пояснения, выводы и комментарии (в текстовых ячейках);
- 2) файл в формате *pdf* (или *html*), полученный путем экспорта (или вывода на печать) ноутбука из п. 1).

Внимание: в названии файлов должна обязательно присутствовать фамилия автора. Безымянные работы проверяться не будут.

Обратите внимание, что все необходимые для выполнения задания методы и программные конструкции рассмотрены в учебных ноутбуках, размещенных в системе LMS. После изучения этих материалов выполнение задания не потребует больших усилий.

Максимальная оценка за выполнение задания вне аудитории – 1 балл. Дополнительные баллы (от 0 до 3) можно будет получить на следующем практическом занятии по результатам тестирования.

Внимание: самостоятельное и вдумчивое выполнение задания серьезно повышает вероятность успешного прохождения теста.

В этой лабораторной работе используется набор данных о продажах домов в Нью-Йорке, уже рассматривавшийся в лабораторной работе 8. [Описание признаков исходного набора данных](#) (напоминание).

Наборы данных по вариантам представлены в csv-файлах; они имеют одну и ту же структуру, соответствующую общему описанию, но отличаются набором записей. Имена файлов: Вариант N, где N – номер варианта.

Задание 1

Реализовать обработку пропущенных значений во всех столбцах набора данных, содержащих пропуски. В процессе работы выполнить следующие действия.

1. Импортировать из файла данные о проданных домах.
2. Вспомнить описание признаков (по ссылке выше); изучить импортированные данные на наличие пропущенных значений.

3. Для каждого признака, имеющего пропущенные значения

- определить тип признака;
- определить процентное соотношение записей с пропусками в общем количестве записей набора данных;
- выполнить исследование, позволяющее сформулировать предположение о механизме формирования пропусков (можно ли считать случайным появление пропущенных значений), привести (написать в текстовых ячейках) подробные комментарии к выполняемым действиям и полученные выводы;
- руководствуясь результатами выполнения предыдущих пунктов, а также априорной информацией о природе признака, обосновать (записать рассуждения в текстовых ячейках) стратегию обработки пропущенных значений;
- реализовать выбранную стратегию и выполнить обработку пропусков.

Замечание 1. При выборе способа восстановления пропущенных значений можно ориентироваться не только на рассмотренные в учебном ноутбуке приемы, но и использовать информацию из других столбцов. Например, можно обратить внимание на взаимосвязи значений в разных столбцах (точный адрес – географические координаты; общая площадь дома – количество ванных и спален, и др.). Следует иметь в виду, что иногда можно точно восстановить пропущенные значения, опираясь на информацию в других столбцах.

Замечание 2. Несмотря на то, что такие столбцы, как географические координаты дома, навряд ли могут быть полезны в качестве входных признаков модели прогнозирования цены, значения в этих столбцах могут использоваться для других целей. Поэтому задача обработки пропусков в этих столбцах может представлять отдельный интерес.

4. Дополнительное задание (+0,25 балла):

- оценить результаты заполнения пропусков, полученные при выполнении этого задания, путем сравнения их с данными из файла лабораторной работы 8; комментарии записать в текстовой ячейке;
- проанализировать возможные причины неудачного восстановления (анализ и полученные выводы записать в текстовых ячейках); предложить альтернативные варианты действий.

Задание 2

Выполнить предварительную обработку предполагаемых входных признаков модели прогнозирования цены дома. В качестве исходных данных использовать набор, полученный в результате выполнения задания 1.

1. Удалить из набора, подлежащего дальнейшей обработке, признаки ADDRESS, STATE, MAIN_ADDRESS, LONG_NAME, FORMATTED_ADDRESS, LATITUDE и LONGITUDE.

2. Удалить также прогнозируемый признак PRICE (он должен обрабатываться отдельно).
3. Подготовить оставшиеся признаки к использованию их в построении модели:
 - для каждого количественного признака обосновать применение наиболее подходящего (с вашей точки зрения) метода масштабирования;
 - реализовать выбранные методы масштабирования;
 - для категориальных признаков обосновать наиболее подходящий метод преобразования категорий в числовые значения;
 - реализовать кодирование категориальных признаков;
 - записать в текстовых ячейках комментарии ко всем выполняемым действиям.