Тема 8. Разведочный анализ данных

Практические задания для самостоятельного выполнения

Задания выполняются по вариантам. Формулировки заданий общие для всех вариантов; набор данных, подлежащих анализу, 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*, 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.

Задание (максимум 3 балла).

Выполнить разведочный анализ данных о продажах домов в Нью-Йорке. Описание признаков исходного набора данных.
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.

В ходе проведения анализа выполнить следующие действия.

- Импортировать из файла данные о проданных домах.
- Изучить описание всех признаков, характеризующих продажи (по ссылке выше).
- В рамках первичного знакомства с данными:
  - вывести несколько записей (для проверки корректности импорта и получения первого представления о данных);
  - изучить признаки на наличие пропущенных значений, типы данных; сопоставить типы столбцов и значения в столбцах с описанием признаков, сделать выводы о корректности имеющихся значений.
- Выполнить исследование одномерных распределений количественных входных признаков BEDS, BATH, PROPERTYSQFT, а также прогнозируемого признака PRICE:
  - для каждого признака найти описательные статистики, асимметрию и эксцесс;
  - визуализировать распределения;
  - проанализировать степень асимметричности, «хвосты», наличие в данных групп, аномальных значений – на основе полученных статистик и визуального оценивания;

- сформулировать предположения о нормальности/отличии от нормального распределения каждого из рассмотренных признаков;
- для признаков, распределение которых было оценено как близкое к нормальному, выполнить визуальную оценку соответствия гистограммы и предполагаемого распределения;
- для непрерывных признаков, распределение которых заметно отличается от нормального вследствие явной асимметрии, проверить предположение о возможной принадлежности к логнормальному распределению (на основе визуальной оценки соответствия гистограммы и теоретического распределения).

- Выполнить исследование на наличие связей между признаками BEDS, BATH и PROPERTYSQFT, а также зависимости между каждым из этих признаков и прогнозируемым признаком PRICE:
  - для каждой пары признаков обосновать применение корреляции Пирсона либо ранговой корреляции;
  - применить (в учебных целях) все три метода корреляционного анализа, проанализировать полученные результаты; сделать выводы;
  - построить парные графики рассеяния (можно использовать seaborn.pairplot(), либо pandas.plotting.scatter_matrix()), соотнести результаты визуализации с результатами корреляционного анализа;
  - сформулировать выводы о возможном наличии/отсутствии зависимостей между изученными признаками.

- Выполнить исследование одномерных распределений категориальных признаков BROKERTITLE, TYPE, LOCALITY и SUBLOCALITY: построить столбцовые диаграммы и изучить распределение категорий, обращая внимание на возможное присутствие малочисленных категорий.
  Пояснение. Малочисленные категории в перспективе обучения предсказательной модели могут рассматриваться как аналоги «тяжелых хвостов»/выбросов для количественных переменных. Наличие таких особенностей может затруднять обучение модели регрессии, приводить к нестабильной работе алгоритма обучения.

- Выполнить исследование на наличие связи между признаками BROKERTITLE и LOCALITY (из-за небольшого объема набора данных – только по наиболее крупным агентствам):
  - отобрать записи, относящиеся к агентствам недвижимости, входящим в топ-3 по общему количеству продаж (в пределах рассматриваемого набора);
  - построить таблицу сопряженности рассматриваемых признаков;
  - оценить правомерность применения критерия «хи-квадрат»; в случае значительного числа клеток с малыми частотами – выполнить объединение малочисленных категорий признака LOCALITY (например, в одну категорию «Others»);
  - применить критерий «хи-квадрат», сделать выводы о наличии связи;

    - вычислить коэффициент Крамера и оценить силу взаимосвязи между признаками (при наличии).

- Выполнить исследование на наличие зависимости между каждым из рассмотренных качественных признаков и прогнозируемым признаком PRICE:
    - выполнить визуализацию в виде диаграмм «ящик с усами»; проанализировать полученные диаграммы, сформулировать предположения о наличии/отсутствии зависимостей;
    - соотнести «ящики с усами» с построенными ранее столбцовыми диаграммами, проанализировать возможность снижения размерности данных путем объединения категорий с близкими значениями прогнозируемого признака;
    - в тех случаях, где это целесообразно, выполнить объединение значений признаков в более крупные категории; построить «ящики с усами» на объединенных категориях.

- По каждому пункту исследования сделать выводы (записать в текстовых ячейках). Привести все необходимые пояснения и комментарии.

- Рассмотреть вопрос о целесообразности использования в предсказательной модели признаков из исходного набора данных, не рассмотренных в данном исследовании. Привести аргументы «за» или «против» необходимости дополнительного изучения этих признаков.

- Дополнительное задание (+0,25 балла):
    - разбить весь имеющийся набор записей на три ценовые группы (низкие, средние и высокие цены);
    - повторить проведенное ранее исследование отдельно для каждой группы; сопоставить описательные статистики, корреляции, диаграммы, полученные для каждой группы; сравнить их с показателями полного набора записей;
    - анализ результатов и полученные выводы описать в текстовых ячейках.

Замечание.
Для более тщательного изучения вопроса о наличии зависимости количественного признака от категорий можно применять дисперсионный анализ – ANOVA (за пределами данного курса). Для желающих познакомиться с методом:

Простое изложение.

Презентация с подробным разбором.

ANOVA в Python.