Конструирование признаков (feature engineering) и подготовка данных

PIERIAN DATA

Конструирование признаков и подготовка данных

- Очень часто, во многих задачах, данные не готовы к машинному обучению – сначала их нужно очистить, а также подготовить более полезные признаки.
- В этом разделе мы возьмём большой набор данных для линейной регрессии, и подготовим его к проекту машинного обучения.

Конструирование признаков и подготовка данных

- Что такое конструирование признаков?
  - Это процесс применения знаний о предметной области, чтобы из сырых данных извлечь полезные признаки, используя техники обработки данных
  - Но как именно это делается?

Конструирование признаков и подготовка данных

- Три основных подхода:
  - Извлечение информации (extract)
  - Комбинирование информации (combine)
  - Преобразование информации (transform)

Конструирование признаков и подготовка данных

- Извлечение информации:
  - Представьте данные о расходах на поездку
  - Для каждой строки есть timestamp:
    - 1990-12-01 09:26:03
  - В таком формате эти данные будет сложно подать на вход алгоритма машинного обучения. Многие алгоритмы принимают числовые данные - float или int

Конструирование признаков и подготовка данных

- Извлечение информации:
  - Вместо даты мы берём отдельную информацию:
    - 1990-12-01 09:26:03
      - Год: 1990
      - Месяц: 12
      - Рабочий день или выходной (0 / 1)
      - День недели: пн (1), вт (2), ср (3) и т.д.

Конструирование признаков и подготовка данных

- Извлечение информации:
  - Более сложные примеры
    - Текстовые документы
      - Длина текста
      - Как часто встречается то или иное ключевое слово

Конструирование признаков и подготовка данных

- Комбинирование информации:
  - Мы уже это делали для полиномиальной регрессии!
  - Вспомните – у нас были бюджеты на рекламу по ТВ, радио и в газетах, и мы добавляли слагаемые с умножением отдельных бюджетов.

Конструирование признаков и подготовка данных

- Комбинирование информации:
  - Также можно комбинировать извлечённые данные
    - Новый признак:
      - вечер рабочего дня (0 / 1)

Конструирование признаков и подготовка данных

- Преобразование информации:
  - Очень часто применяется для текстовых данных
  - Многие алгоритмы не могут работать с текстовыми данными (нельзя умножить слово “красный” на числовой коэффициент)

Конструирование признаков и подготовка данных

- Преобразование информации:
  - Категориальные данные часто приходят в текстовом виде
  - Например, в наборе данных может быть указана страна пользователя как строковое значение (USA, UK, MEX, ...)

Конструирование признаков и подготовка данных

- Преобразование информации:
  - Категориальные данные часто приходят в текстовом виде
  - Например, в наборе данных может быть указана страна пользователя как строковое значение (USA, UK, MEX, ...)
  - Здесь можно применить два подхода:
    - Кодировка числами
    - Кодировка одного значения (dummy-переменные)

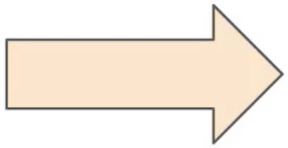
Конструирование признаков и подготовка данных

- Кодировка числами (integer encoding):
  - Назначаем категориям номера – 0, 1, 2, 3 и т.д.

Конструирование признаков и подготовка данных

- Кодировка числами (integer encoding):
  - Назначаем категориям номера – 0, 1, 2, 3 и т.д.

Country
USA
MEX
CAN
USA

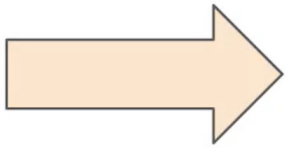


Country
1
2
3
1

Конструирование признаков и подготовка данных

- Кодировка числами (integer encoding):
  - Назначаем категориям номера – 0, 1, 2, 3 и т.д.

Country
USA
MEX
CAN
USA



Country
1
2
3
1

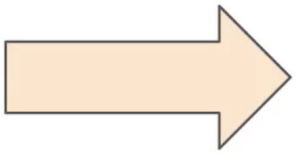
Возможная проблема: неявно упорядочиваем значения

PIERIAN DATA

Конструирование признаков и подготовка данных

- Кодировка числами (integer encoding):
  - Иногда упорядоченные номера имеют смысл:

Spice Level
Mild
Hot
Fire
Mild



Spice Level
1
2
3
1

Конструирование признаков и подготовка данных

- Кодировка числами (integer encoding):
  - Плюсы:
    - Легко сделать и понять
    - Не увеличивает количество признаков
  - Минусы:
    - Добавляет упорядоченность между категориями

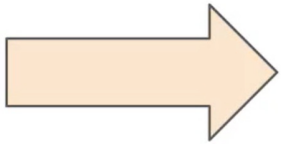
Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Для каждой категории создаём отдельный признак (dummy-переменную) со значением 0 или 1

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Для каждой категории создаём отдельный признак (dummy-переменную) со значением 0 или 1

Country
USA
MEX
CAN
USA

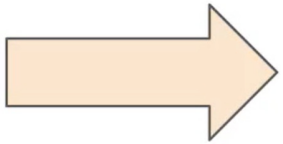


USA	MEX	CAN
1	0	0
0	1	0
0	0	1
1	0	0

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Для каждой категории создаём отдельный признак (dummy-переменную) со значением 0 или 1

Country
USA
MEX
CAN
USA



Значения не упорядочены

USA	MEX	CAN
1	0	0
0	1	0
0	0	1
1	0	0

PIERIAN DATA

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Большое количество дополнительных признаков
  - Имеет смысл брать крупные категории, например регионы вместо отдельных стран

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Большое количество дополнительных признаков
  - Имеет смысл брать крупные категории, например регионы вместо отдельных стран
  - В Pandas для этих целей есть функции .map() и .apply()
  - Может понадобится время на выбор оптимального уровня категорий.

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Также нужно помнить про “ловушку dummy-переменных”, математически известную как мульти-коллинеарность.
  - Конвертация в dummy-переменные может приводить к дублированию признаков.
  - Давайте рассмотрим простейший пример...

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Рассмотрим бинарную переменную (только два значения)

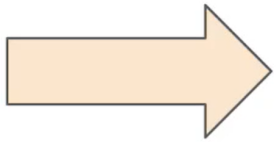
Vertical Direction
UP
DOWN
UP
DOWN

PIERIAN DATA

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Рассмотрим бинарную переменную (только два значения)

Vertical Direction
UP
DOWN
UP
DOWN

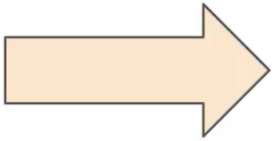


UP	DOWN
1	0
0	1
1	0
0	1

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Рассмотрим бинарную переменную (только два значения)
  - Две новые колонки дублируют друг друга (с инверсией)

Vertical Direction
UP
DOWN
UP
DOWN

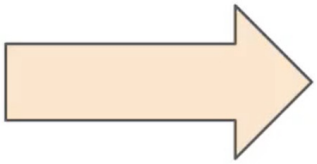


UP	DOWN
1	0
0	1
1	0
0	1

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Это применимо и в случае более двух категорий:

Country
USA
MEX
CAN
USA



USA	MEX
1	0
0	1
0	0
1	0

Конструирование признаков и подготовка данных

- Кодировка одного значения (one hot encoding):
  - Плюсы
    - Не добавляется упорядоченность категорий
  - Минусы
    - Добавляется много дополнительных признаков и коэффициентов
    - “Ловушка dummy-переменных”
    - Сложнее добавлять новые категории

Конструирование признаков и подготовка данных

- В этом разделе курса мы рассмотрим:
  - Выбросы в данных (outliers)
  - Отсутствующие данные (missing data)
  - Категориальные данные
- Не все эти задачи являются “конструированием признаков”, часть из них это “очистка данных”.

Конструирование признаков и подготовка данных

- Имейте ввиду, что в общем случае конструирование признаков зависит от данных и контекста.
- Не существует единого решения на все случаи жизни!

Работа с выбросами (outliers)

PIERIAN DATA

Работа с выбросами в данных (outliers)

- Часто в данных есть несколько точек, экстремально отличающиеся от всех других точек.
- Зачастую лучше просто удалить эти точки из набора данных, чтобы получить более удачную модель

Работа с выбросами в данных (outliers)

- Какое значение считать выбросом (outlier)?
  - Диапазоны и лимиты
  - Процент строк данных
  - И то, и другое очень зависит от конкретной ситуации!

Работа с выбросами в данных (outliers)

- Какое значение считать выбросом (outlier)?
  - Диапазоны и лимиты
    - Мы должны решить, что считать выбросом, на основе некоторой методологии:
      - Интерквартильный диапазон
      - Среднеквадратичное отклонение
      - Визуализация или знания о природе признака

PIERIAN DATA

Работа с выбросами в данных (outliers)

- Какое значение считать выбросом (outlier)?
  - Процент строк данных
    - Если большой процент строк выглядит как выбросы, то это просто широкий диапазон возможных значений признака
    - Процент выбросов не должен превышать максимум нескольких процентов

Работа с выбросами в данных (outliers)

- Какое значение считать выбросом (outlier)?
  - Полезно визуализировать данные, чтобы увидеть точки-выбросы
  - Имейте ввиду, что это может привести к погрешностям в будущей модели (например, модель не подходит для домов дороже $10 миллионов).

Работа с выбросами в данных (outliers)

- Имейте ввиду, что не существует на 100% корректной методики определения точек-выбросов для всех ситуаций.
- Давайте поищем выбросы в наборе данных Ames!

Отсутствующие данные (missing data)

Часть 1: каких данных не хватает

PIERIAN DATA

Отсутствующие данные (missing data)

- Пожалуйста убедитесь, что Вы посмотрели лекции “Отсутствующие данные” в разделе Pandas перед тем, как продолжить просмотр этой лекции.

Отсутствующие данные (missing data)

- Пожалуйста убедитесь, что Вы посмотрели лекции “Отсутствующие данные” в разделе Pandas перед тем, как продолжить просмотр этой лекции.
- Мы будем работать с набором данных Ames. В этой лекции мы выясним, каких данных не хватает – сколько и каких именно.

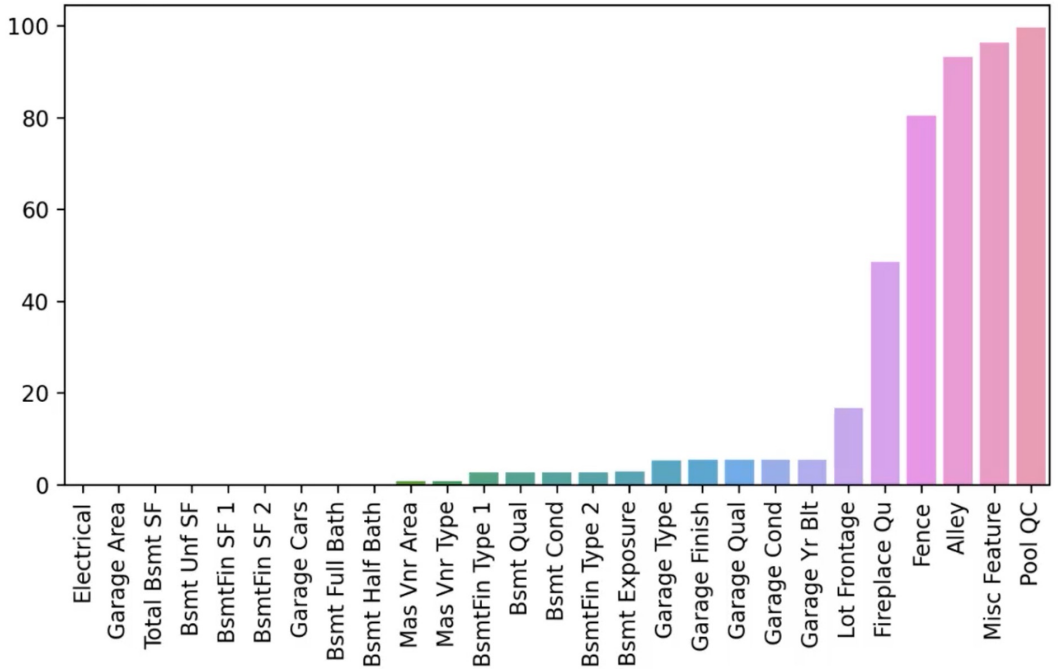
Отсутствующие данные (missing data)

Часть 2: заполнение данных по строкам

PIERIAN DATA

Отсутствующие данные (missing data)

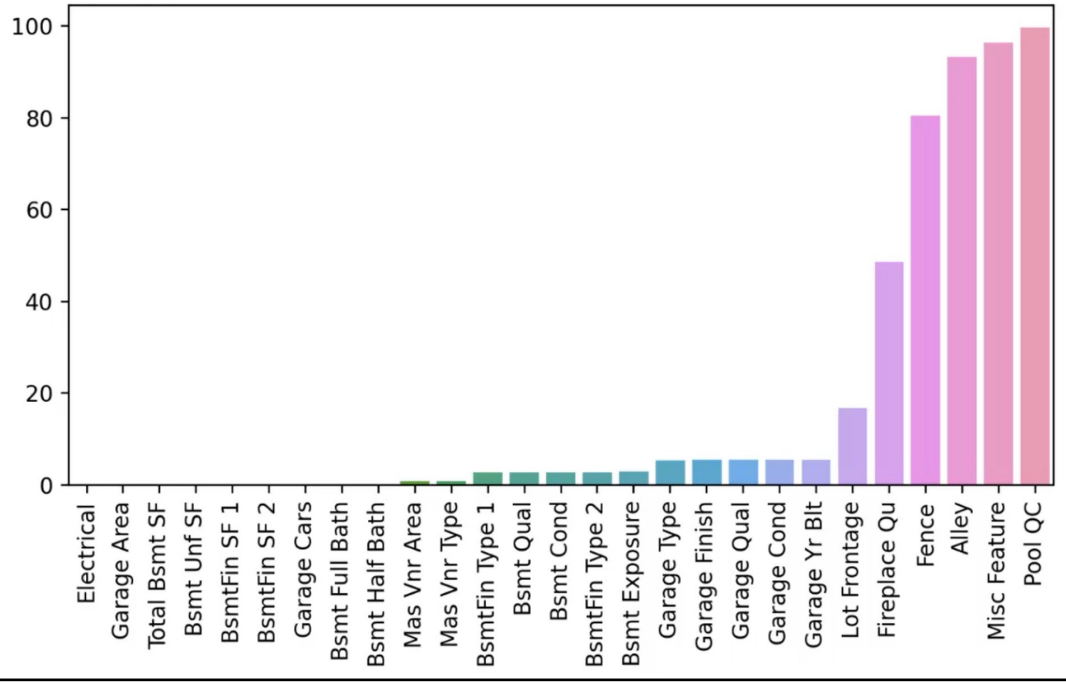
- Ранее мы посчитали процент отсутствия данных для каждой колонки с признаками.



PIERIAN DATA

Отсутствующие данные (missing data)

- Для начала рассмотрим те колонки, где процент маленький.



PIERIAN DATA

Отсутствующие данные (missing data)

- Если данных нет всего в нескольких строках, то можно:
  - либо удалить эти несколько строк
  - либо заполнить их каким-то средним значением, учитывая наши знания об этой колонке.
- Давайте перейдём в блокнот и посмотрим

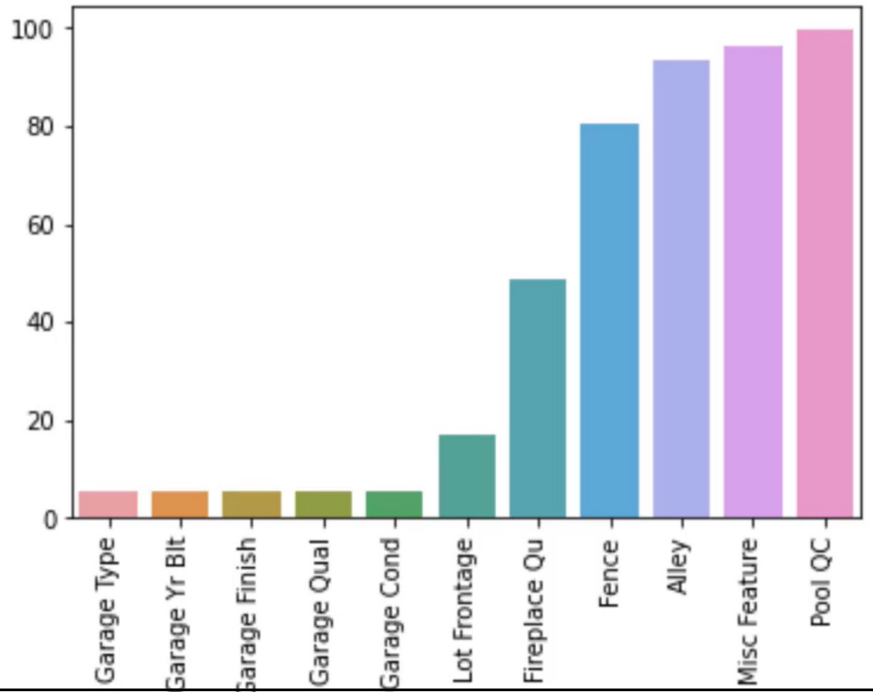
Отсутствующие данные (missing data)

Часть 3: заполнение данных по колонкам

PIERIAN DATA

Отсутствующие данные (missing data)

- Возьмём колонки, где процент отсутствия данных выше порогового значения 1%



PIERIAN DATA

Отсутствующие данные (missing data)

- Два подхода:
  - Заполнить колонки некоторыми значениями
  - Удалить такие колонки
- Рассмотрим плюсы и минусы каждого из подходов...

Отсутствующие данные (missing data)

- Подход “удалить такие колонки”:
  - Очень просто сделать
  - Не нужно больше беспокоиться об этих признаках
  - Но мы можем потерять признак с важными данными
  - Удалять колонки имеет смысл, когда много строк имеют значения NaN

Отсутствующие данные (missing data)

- Подход “заполнить каким-то значением”:
  - Потенциально меняем истинность исходных данных
  - Мы сами должны выбрать способ, какое значение записать
  - Нужно будет применять трансформации для всех будущих данных

Отсутствующие данные (missing data)

- Подход “заполнить каким-то значением”:
  - Простой случай
    - Заменить значения NaN на нули, если по факту неопределённые значения это нулевые значения
  - Сложные случаи
    - Применяем статистические методы с использованием других колонок, чтобы заполнить значения NaN

Отсутствующие данные (missing data)

- Подход “заполнить каким-то значением”:
  - Статистический метод
    - В наборе данных не хватает информации о возрасте
    - Мы можем использовать информацию о текущей работе или образовании, чтобы по ним заполнить возраст (например, если сейчас человек учится в колледже, то записать возраст 20 лет)

Отсутствующие данные(missing data)

- Давайте рассмотрим оба метода!
  - Важное замечание:
    - В наборе данных Ames большинство значений NaN можно корректно заменить значениями 0. Но мы рассмотрим разные методы, чтобы изучить их!

Работа с категориальными данными

PIERIAN DATA

Работа с категориальными данными

- Мы сразу перейдем к работе в блокноте.
- Убедитесь, что Вы полностью посмотрели первую лекцию этого раздела – в ней мы обсуждаем dummy-переменные и технику “one hot encoding”.