

Машинное Обучение

Конструирование признаков (Feature Engineering) и подготовка данных

Конструирование признаков и подготовка данных

- Очень часто данные не готовы к машинному обучению, сначала очистить, а так же подготовить более полезные признаки.
- Наша с вами задача взять большой набор данных и подготовить его к МО.

Конструирование признаков

- Это процесс применения знаний о предметной области, чтобы из сырых данных извлечь полезные признаки, используя техники обработки данных.

Конструирование признаков

Три основных подхода:

- Извлечение информации (extract)
- Комбинирование информации (combine)
- Преобразование информации (transform)

Конструирование признаков

Извлечение информации (extract):

- 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

Извлечение информации (extract):

- 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

Извлечение информации (extract):

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

Конструирование признаков

Комбинирование информации:

Создание новых признаков:

- Вечер рабочего дня
- Обеденное время

Конструирование признаков

Преобразование данных:

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

Конструирование признаков

Преобразование данных:

- 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

Кодировка числами:

- Назначаем категориям номера – 0, 1, 2, 3 и т.д.

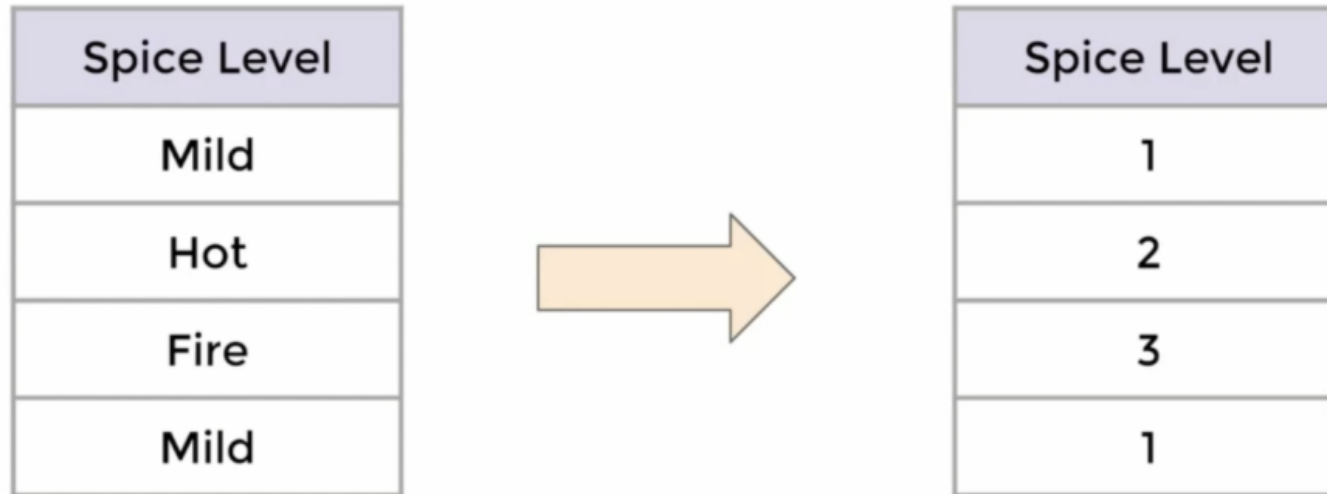
Country	Country
USA	1
MEX	2
CAN	3
USA	1

Возможная проблема: неявное упорядочивание значений

Конструирование признаков

Кодировка числами:

- Иногда упорядоченные номера имеют смысл:



Возможная проблема: неявно говорим, что Fire в 2 раза острее, чем hot

Конструирование признаков

Кодировка числами:

- Плюсы:
 - Легко сделать и понять
 - Не увеличивает количество признаков
- Минусы:
 - Добавляет упорядоченность между категориями

Конструирование признаков

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

- + признаки независимы друг от друга
- стало 3 переменных

Конструирование признаков

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

Кодировка одного значения (one hot encoding):

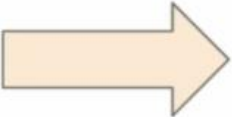
- Также нужно помнить про “ловушку dummy-переменных”, математически известную как мульти-коллинеарность.
- Конвертация в dummy-переменные может приводить к дублированию признаков.
- Давайте рассмотрим простейший пример...

Конструирование признаков

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	

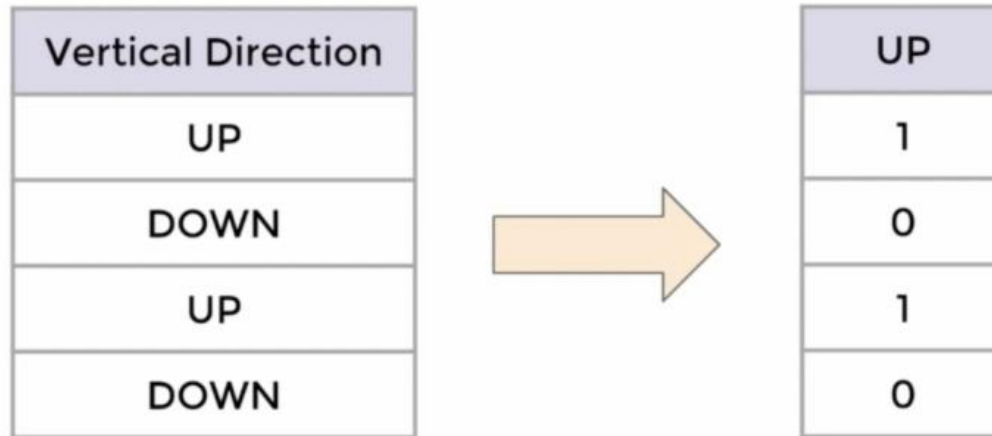


UP	DOWN
1	0
0	1
1	0
0	1

Конструирование признаков

Кодировка одного значения (one hot encoding):

- Рассмотрим бинарную переменную (только два значения)
- Две новые колонки дублируют друг друга (с инверсией)



Конструирование признаков

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

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

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

Предварительная обработка данных

Процесс подготовки данных для использования в модели машинного обучения.

Может включать в себя широкий спектр задач:

- очистка и фильтрация данных – исправление или удаление отсутствующих значений, дубликатов и выбросов из набора данных
- преобразование данных, например преобразование категориальных переменных в числовые переменные или уменьшение размерности набора данных
- масштабирование или нормализацию данных

Подготовка данных

- Работа с выбросами в данных (outliers)
- Работа с отсутствующими данными (missing data) - Часть 1 -
Оценка ситуации
- Работа с отсутствующими данными (missing data) - Часть 2 -
Работа по строкам
- Работа с отсутствующими данными (missing data) - Часть 3 -
Работа по колонкам
- Работа с категориальными переменными

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!

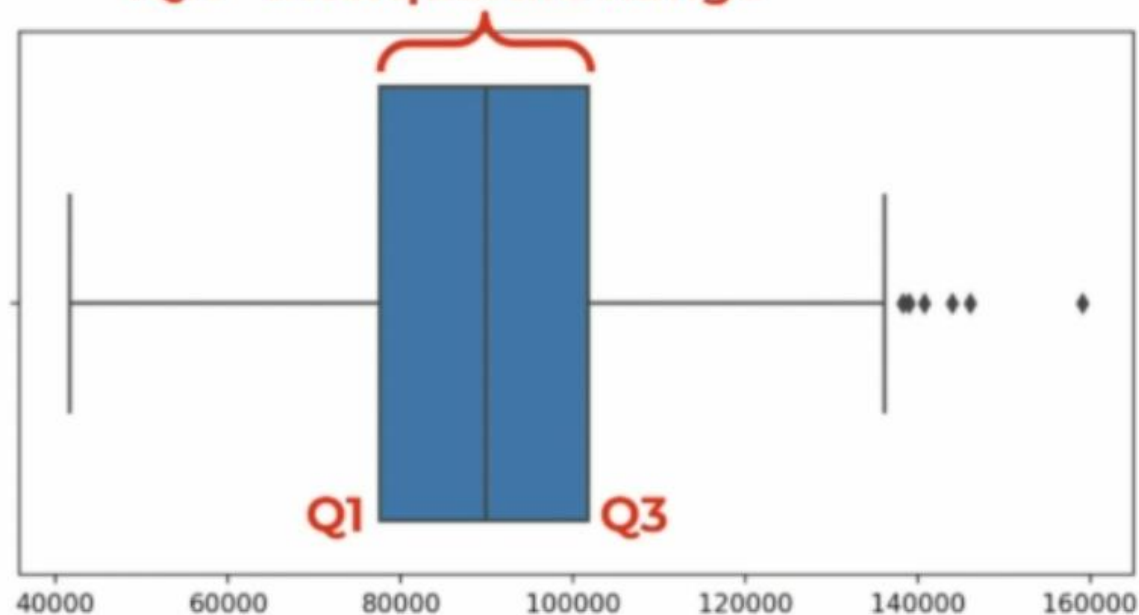
Работа с выбросами в данных (outliers)

- 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

Работа с выбросами в данных (outliers)

Интерквартильный размах – от Q1 до Q3. Это 50% точек.

IQR - Interquartile Range



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

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!

Домашнее задание

- Найти любой массив данных или сгенерировать.
- Произвести подготовку данных на этом наборе данных.

Где взять наборы данных?

- UCI Machine Learning Repository - это один из самых популярных репозиторий наборов данных для машинного обучения. Здесь вы можете найти большое количество наборов данных, которые могут использоваться для обучения моделей машинного обучения и работы с предварительной обработкой данных.
- Kaggle - это платформа для соревнований по машинному обучению, которая также предлагает наборы данных для обучения моделей. Многие из них уже содержат некоторую предварительную обработку данных, но все же могут быть полезными для выполнения лабораторной работы.
- OpenML - это онлайн-платформа для обмена наборами данных и экспериментами в области машинного обучения. Здесь вы можете найти наборы данных, которые могут использоваться для обучения моделей и выполнения лабораторных работ.
- Google Dataset Search - это поисковая система Google для нахождения наборов данных. Здесь вы можете найти наборы данных, которые могут быть полезны для выполнения лабораторной работы.
- Scikit-learn - библиотека Scikit-learn также предоставляет несколько наборов данных, которые могут использоваться для обучения моделей и выполнения лабораторных работ. Вы можете найти их в разделе "datasets" на официальном сайте библиотеки.

Работа с отсутствующими данными (missing data)

Отсутствие данных может возникнуть по различным причинам, таким как повреждение данных, потеря данных, человеческая ошибка или просто потому, что определенная информация не была собрана.

В любом случае отсутствующие данные могут оказать существенное влияние на производительность моделей машинного обучения.

Работа с отсутствующими данными (missing data). Три типа отсутствующих данных

Отсутствует полностью случайным образом (MCAR): когда отсутствующие данные не связаны с какой-либо другой переменной в наборе данных, это называется MCAR. Например, если данные были потеряны во время передачи, это MCAR.

Отсутствует случайно (MAR): когда отсутствующие данные связаны с какой-то другой переменной в наборе данных, он называется MAR. Например, если участник опроса предпочитает не отвечать на конкретный вопрос, основанный на их полу, отсутствующие данные связаны с гендерной переменной.

Отсутствует не случайно (MNAR): когда отсутствующие данные связаны с самим недостающим значением, он называется MNAR. Например, если данные не были собраны, потому что участник не хотел раскрывать свой доход, отсутствующие данные связаны с переменной дохода.

Работа с отсутствующими данными (missing data). Методы удаления

- **Удаление строк со слишком большим количеством пропущенных значений столбцов.**
- **Удаление столбцов со слишком большим количеством пропущенных значений.**

Работа с отсутствующими данными (missing data). **Методы вменения**

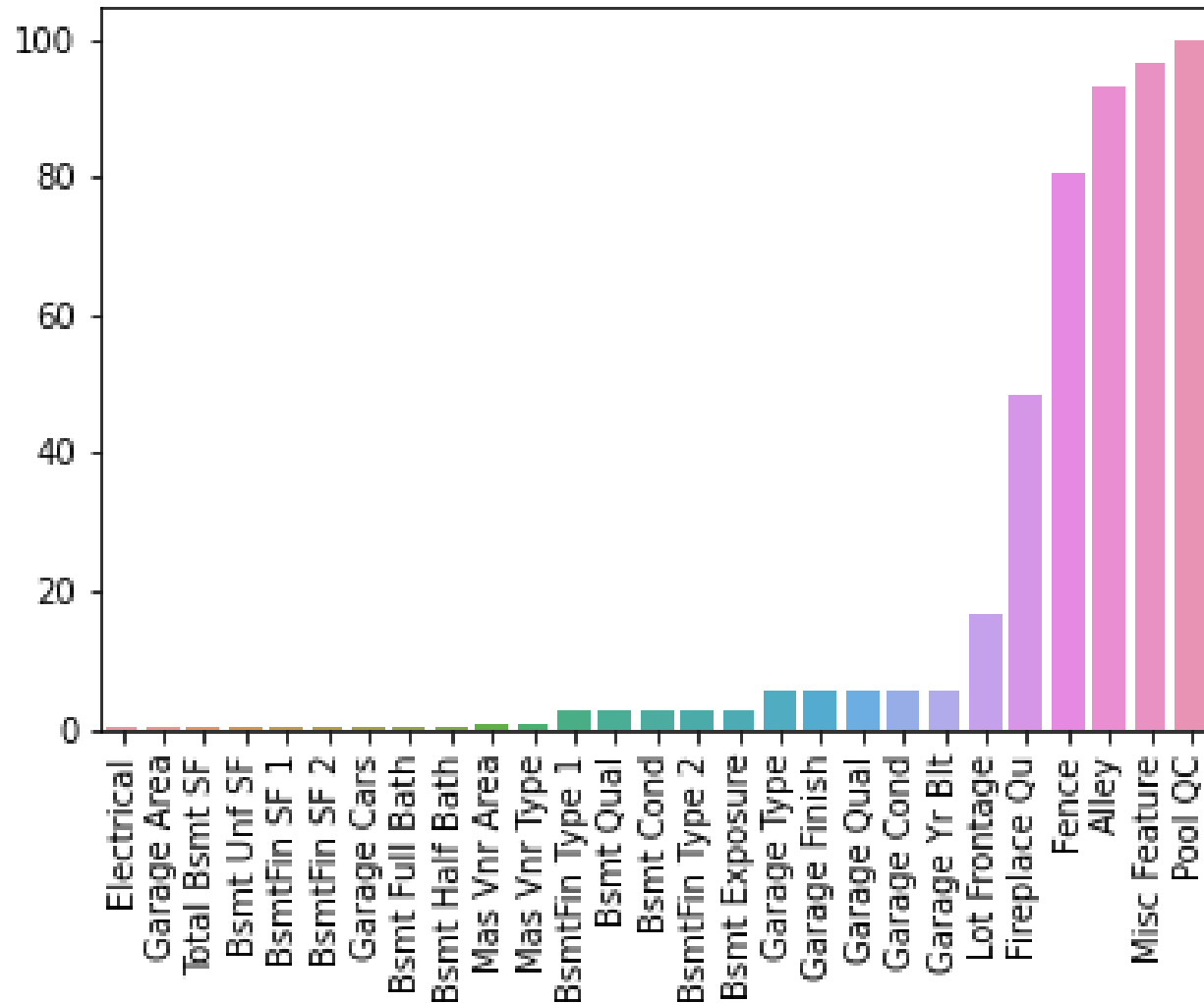
- **Постоянное / среднее / медианное вменение.**
- **Вменение регрессии.**
- **Множественное вменение.**

Работа с отсутствующими данными (missing data). Работа со строками.

Для начала рассмотрим те колонки (признаки), где процент маленький.

Если данных нет всего в нескольких строках, то можно:

- Либо удалить эти несколько строк
- Либо заполнить их каким-то средним значением, учитывая ниши знания об этой колонке.



Работа с отсутствующими данными (missing data). Работа со строками.

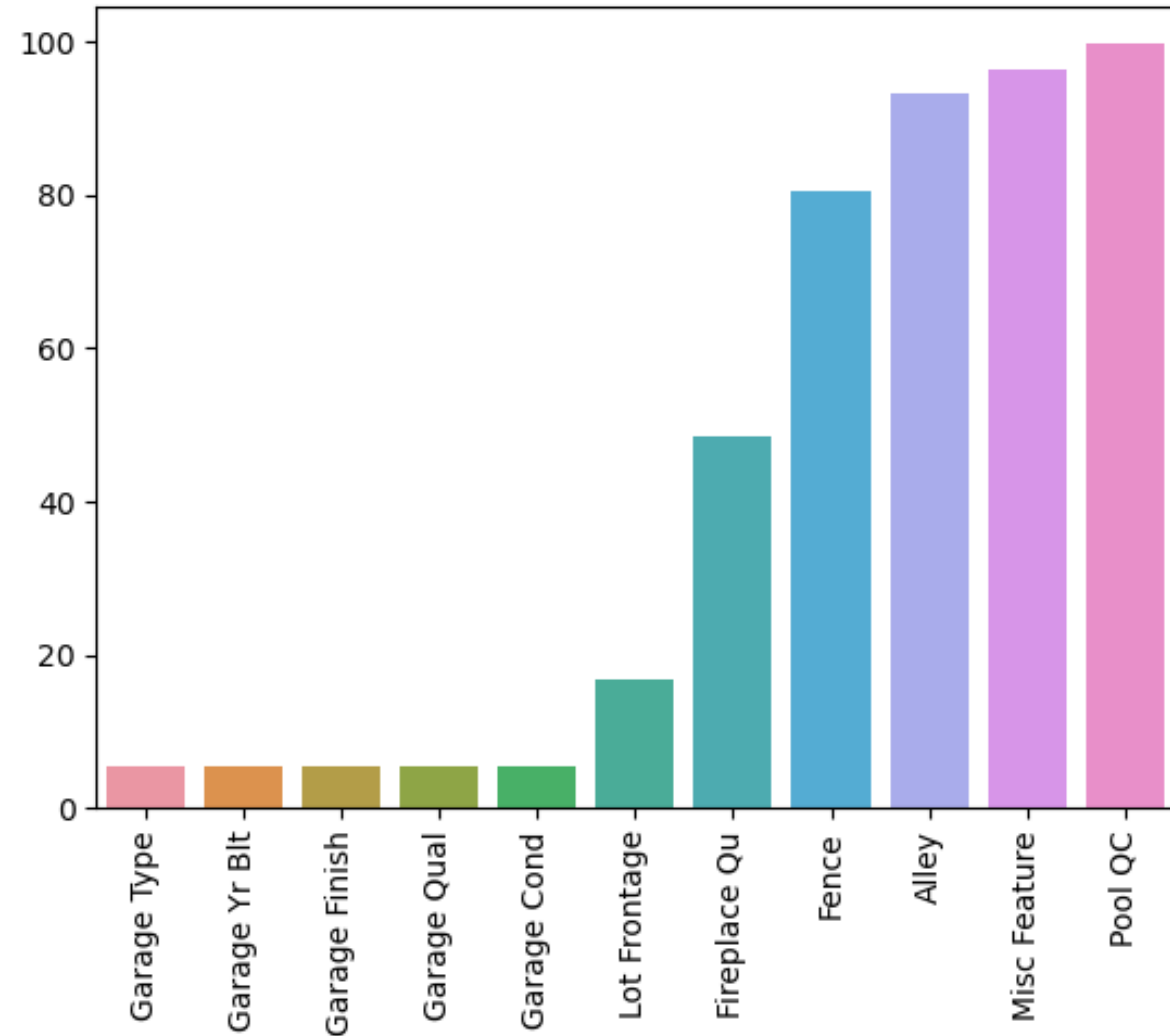
Работа в лабораторной

Работа с отсутствующими данными (missing data). Работа с колонками.

Возьмем колонки, где процент отсутствия данных выше порогового значения 1%

Подходы:

- Заполнить колонки некоторыми значениями
- Удалить такие колонки



Работа с отсутствующими данными (missing data). Работа с колонками.

Удалить такие колонки:

- Очень просто сделать
- Не нужно больше беспокоиться об этих признаках
- Можем потерять признак с важными данными
- Удалять колонки имеет смысл, когда много колонок имеют значение NaN

Заполнить колонки некоторыми значениями:

- Потенциально меняем истинность исходных данных
- Мы сами должны выбрать способ, какое значение записать
- Нужно будет применять трансформации для всех будущих данных

Работа с отсутствующими данными (missing data). Работа с колонками.

Заполнить колонки некоторыми значениями:

- Простой случай
 - Заменить значение NaN на нули, если по факту неопределенные значения это нулевые значения.
- Сложные случаи
 - Применяем статистические методы с использованием других колонок, чтобы заполнить значение NaN

Работа с отсутствующими данными (missing data). Работа с колонками.

Заполнить колонки некоторыми значениями:

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)

Работа с отсутствующими данными (missing data). Работа с колонками.

Лабораторная

Домашнее задание

- Найти любой массив данных с категориальными и цифровыми данными.
- Произвести подготовку данных на этом наборе данных.

Где взять наборы данных?

- UCI Machine Learning Repository - это один из самых популярных репозиторий наборов данных для машинного обучения. Здесь вы можете найти большое количество наборов данных, которые могут использоваться для обучения моделей машинного обучения и работы с предварительной обработкой данных.
- Kaggle - это платформа для соревнований по машинному обучению, которая также предлагает наборы данных для обучения моделей. Многие из них уже содержат некоторую предварительную обработку данных, но все же могут быть полезными для выполнения лабораторной работы.
- OpenML - это онлайн-платформа для обмена наборами данных и экспериментами в области машинного обучения. Здесь вы можете найти наборы данных, которые могут использоваться для обучения моделей и выполнения лабораторных работ.
- Google Dataset Search - это поисковая система Google для нахождения наборов данных. Здесь вы можете найти наборы данных, которые могут быть полезны для выполнения лабораторной работы.
- Scikit-learn - библиотека Scikit-learn также предоставляет несколько наборов данных, которые могут использоваться для обучения моделей и выполнения лабораторных работ. Вы можете найти их в разделе "datasets" на официальном сайте библиотеки.