## ML4.1. Предварительная обработка данных.

### Цель работы

Познакомиться с основными приемами предварительной обработки и очистки данных.

### Задания для выполнения

1. Загрузите и прочитайте в pandas следующий [датасет](https://drive.google.com/file/d/1bGqLj75tOpURPG9awXynfdvJxNXedlPu/view?usp=sharing):

<https://raw.githubusercontent.com/narendramall/Predicting_Price_of_Pre_Owned_Cars/master/cars_sampled.csv>

1. Обратите внимание, что датасет собран «криво», поэтому прежде чем он «правильно» загрузится (рис.1), необходимо его обработать.

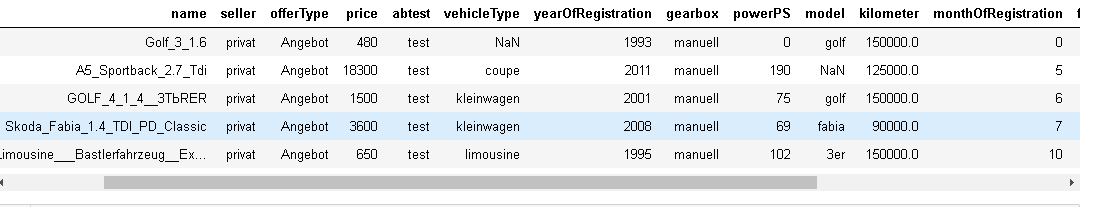


Рисунок 1. Правильно загруженный датасет

1. Сделайте описательную статистику: размер, типы переменных, пустые значения, уникальные имена и т.д.
2. Поработайте с типами переменных. Все числовые значения переведите в формат int.
3. Поработайте со столбцом price. Постройте гистограмму, найдите выбросы (ящиковая диаграмма), удалите аномальные значения, постарайтесь привести к нормальному распределению.
4. Аналогично поработайте и с другими «важными для целевой функции price» параметрами (powerPS, yearOfRegistration, kilometer и т.д). «Важность» докажите через коэффициент корреляции. Сделайте визуализацию.
5. Постройте ящиковые диаграммы зависимости gearbox, fuelType, vehicleType, notRepairedDamage от price. Сделайте выводы. Можно ли убрать значения с малой частотой? Повлияет ли это на дальнейшее исследование?
6. Найдите пропуски. Удалите столбцы, которые содержат большое количество пропусков. В остальных случаях постарайтесь восстановить пропущенные значения.
7. Сохраните полученный датасет и выведете его.

### Методические указания

Реальные данные, используемые для машинного обучения на практике могут происходить из различных источников и содержать в себе ошибки. Эти ошибки - это следствие человеческого фактора, багов в программах обработки данных, неполноты данных в источниках, и множества других причин. Для эффективного использования данных такие ошибки нужно исправить.

Для иллюстрации наиболее распространенных ошибок в данных давайте возьмем фрагмент датасета из задачи про кредитный скоринг:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 100000000 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| А | 28 | СПб | 80000 | B | 0 |
| М | 23 |  | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

В этом небольшом участке видно сразу несколько несостыковок. Например, сразу же в первом столбце присутствует аномальное значение:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 100000000 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| А | 28 | СПб | 80000 | B | 0 |
| М | 23 |  | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Далее, во втором столбце 10 - это тоже, скорее всего, неправильное значение. Вряд ли десятилетний ребенок обращался в банк за кредитом:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 100000000 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| А | 28 | СПб | 80000 | B | 0 |
| М | 23 |  | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Переходим к третьему столбцу. Здесь все еще очевиднее. В одной из строк значение просто отсутствует. Наверное, его просто забыли внести при заполнении форм, или оно потерялось в процессе преобразования. Для большинства алгоритмов машинного обучения отсутствующие значения недопустимы. Поэтому этот пробел надо будет чем-то заполнить:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 100000000 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| А | 28 | СПб | 80000 | B | 0 |
| М | 23 | ? | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

В четвертом столбце тоже не все чисто. Присутствует аномально высокое значение. Непонятно даже, откуда оно взялось. Может быть это ошибка, а может быть, какое-то специальное значение (типа MAXINT). Вообще, это может быть просто выбросом, то есть корректным значением, но нехарактерно далеким от среднего. Ведь миллиардеры тоже могут брать кредиты.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 100000000 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| А | 28 | СПб | 80000 | B | 0 |
| М | 23 |  | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Если присмотреться, в этом же столбце есть еще одно подозрительное место. Обратите внимание, что все значения здесь приведены в десятках тысяч. Однако одно конкретное значение приведено с нехарактерной точностью:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 100000000 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| А | 28 | СПб | 80000 | B | 0 |
| М | 23 |  | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Перейдем к последнему признаку. Здесь вообще без пояснения шкалы непонятно, что эти буквы означают. Самый главный вопрос: можно ли их сравнивать, то есть установлено ли у этого признака отношение порядка. Ведь если это уровень образования (начальный, средний, высший), то их можно представить как числа. А если это направление (техническое, гуманитарное, естественное), то нужно к этому признаку относиться как к категориальному.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 100000000 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| А | 28 | СПб | 80000 | B | 0 |
| М | 23 |  | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Итого в нашем фрагменте несколько сомнительных значений, которые лучше удалить и заполнять более правдоподобными:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | ? | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| ? | 28 | СПб | 80000 | B | 0 |
| М | 23 | ? | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Отсутствующие значения в числовых признаках чаще всего заменяют на среднее значение по этому признаку. Это дает наименьшее отклонение от изначального распределения признака:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 49994.6 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| ? | 28 | СПб | 80000 | B | 0 |
| М | 23 | ? | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Хотя иногда используют не среднее арифметическое, а медианное значение:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | 40000 | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| ? | 28 | СПб | 80000 | B | 0 |
| М | 23 | ? | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Пропуски в категориальных признаках можно заменять на наиболее популярное (модальное) значение:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | ? | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| М | 28 | СПб | 80000 | B | 0 |
| М | 23 | Мск | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Хотя, если пропусков довольно много, то такая замена сильно сместит распределение к центру. Поэтому довольно часто пропуски в категориальных переменных заполняют специальным значением, обозначающим отсутствие значения. Потом, когда мы будем кодировать этот признак, это приведет к созданию еще одной фиктивной переменной. Такой способ сохраняет наибольшее количество информации из исходного набора данных. Он даже может учитывать, если сам пропуск несет какую-то смысловую нагрузку.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | ? | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| Пр | 28 | СПб | 80000 | B | 0 |
| М | 23 | Пр | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

Еще один способ заполнения пропусков - случайными значениями из распределения данного признака:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Пол | Возраст | Город | Доход | Образование | Вернул? |
| М | 25 | Мск | 70000 | A | 1 |
| Ж | 31 | Мск | ? | C | 1 |
| М | 10 | Мск | 40000 | C | 0 |
| Ж | 28 | СПб | 80000 | B | 0 |
| М | 23 | Екб | 40000 | A | 1 |
| Ж | 30 | Екб | 19973 | B | 1 |

### Контрольные вопросы

1. Какие основные проблемы в датасете из лабораторной вы обнаружили?
2. Каковы основные методы заполнения пропущенных значений в данных?
3. Какие существуют методы обнаружения аномальных значений в числовых признаках?

### Дополнительные задания

1. Постройте модель регрессии для цены автомобиля. Обучите ее на данных, полученных из исходного датасета с применением разных методов заполнения пропущенных значений. Сделайте вывод об эффективности разных методов.