Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Машинное обучение»

Студент: Шорников П. С.
Группа: М8О-407Б-17
Дата:
Оценка:
Подпись:

Москва, 2021

Постановка задачи

Необходимо сформировать два набора данных для приложений машинного обучения. **Первый датасет** должен представлять из себя табличный набор данных для задачи классификации. **Второй датасет** должен быть отличен от первого, и может представлять из себя набор изображений, корпус документов, другой табличный датасет или датасет из соревнования Kaggle, предназначенный для решения интересующей вас задачи машинного обучения.

Необходимо провести анализ обоих наборов данных, поставить решаемую вами задачу, определить признаки необходимые для решения задачи, в случае необходимости заняться генерацией новых признаков, устранением проблем в данных, визуализировать распределение и зависимость целевого признака от выбранных признаков.

- Логистическая регрессия
- KNN
- SVM
- Дерево Решений

1 Выбранные датасеты

Price of Moscow apartments
(`https://www.kaggle.com/hugoncosta/price-of-flats-in-moscow`)
Mobile Price Classification
(`https://www.kaggle.com/iabhishekofficial/mobile-price-classification`)

2 Price of Moscow apartments

1 Описание входных данных

- Price – цена в тысячах долларов
- Totsp – общая площадь
- Livesp – жилая площадь
- kitsp – площадь кухни

- dist – расстояние от центра
- metrdist – расстояние до метро в минутах
- walk – 1 – пешком от метро, 0 – на транспорте
- brick - 1 – кирпичный, монолит ж/б, 0 – другой
- floor – 1 – этаж кроме первого и последнего, иначе – 0.

1.1 Анализ данных

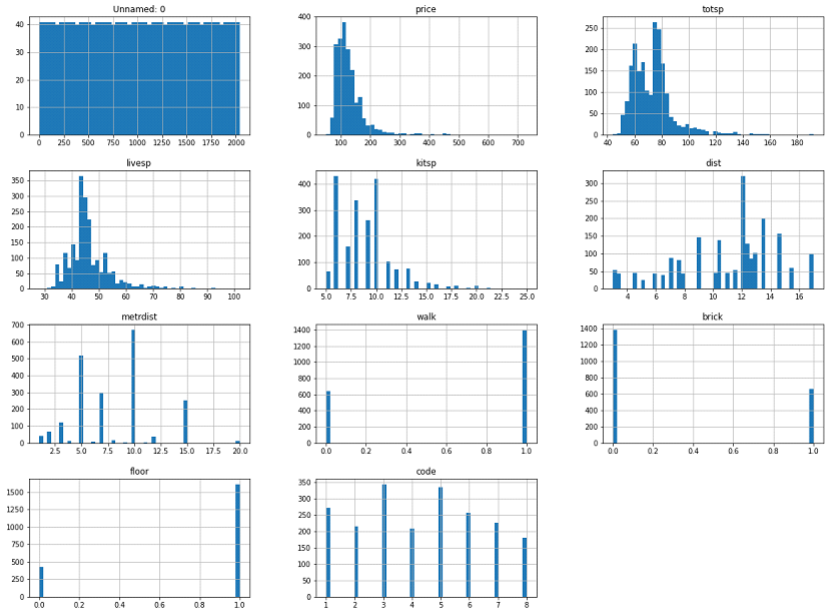
Типы признаков

- Количественные признаки: все

Размер

- Строки: 2040
- Столбцы: 10

Распределение признаков с числовыми полями



3 Mobile Price Classification

1 Описание входных данных

- city – город
- car_maker – марка
- model – модель
- year – год производства
- condition– состояние
- kilometers – пробег
- transmission– трансмиссия
- fuel – топливо
- color – цвет
- pay_method – метод оплаты
- price – цена

1.1 Анализ данных

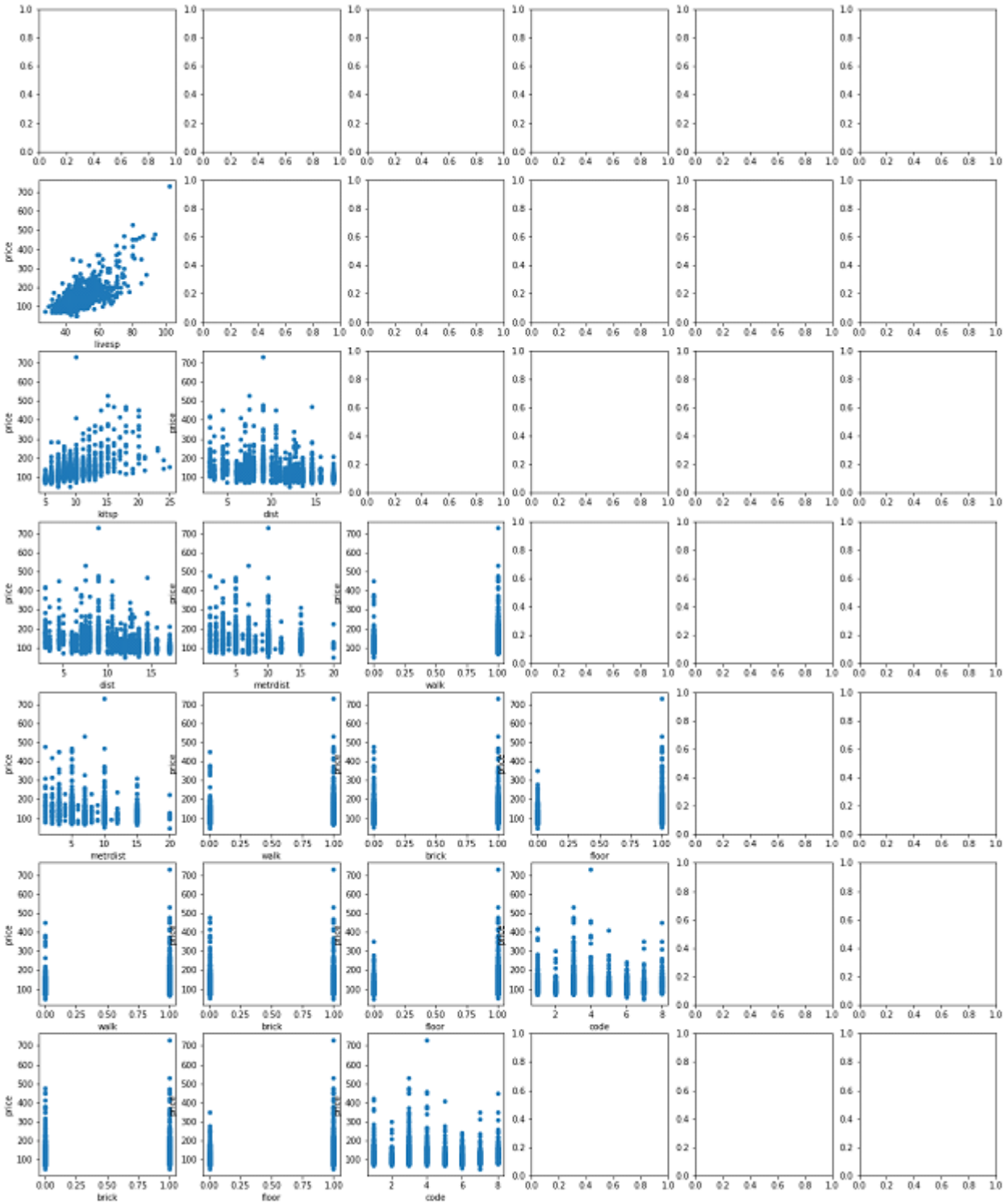
Типы признаков

- Категориальные: city, car_maker, model, condition, transmission, fuel, color, pay_method
- Численные значения: price, kilometers
- Дата: year
- Иследуемое значение: price

Размер

- Строки: 560
- Столбцы: 11

Распределение признаков с числовыми полями



2 Решаемая задача

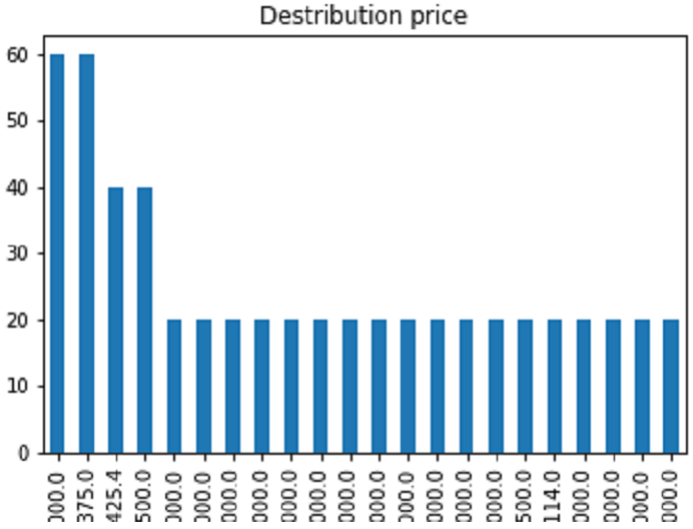
Классифицировать **price**.

2.1 Работа с категориальными признаками

Для оцифровывания категориальных признаков я пользовался lable encoder. Данный метод каждому из уникальных значений в текущем признаке присваивает свою метрику.

3 Визуализация

Распределение по кластерам.



4 Вывод

В ходе лабораторной работы были проанализированы два датасета. Для каждого из них были подготовлены для поставленной задачи данные. Также было показано, как распределение признака, который предстоит исследовать, так и его зависимость от других признаков.