Московский Авиационный Институт

(Национальный Исследовательский Университет)

Отчет по лабораторной работе

по курсу "Искусственый интеллект"

Студент: Ваньков Д. А.

Группа: М80-307Б-17

Преподаватель:

Москва, 2020

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. В отчете описать все проблемы, с которыми вы столкнулись и выбранные подходы к их решению.

Выбранные датасеты:

Melbourne Housing Snapshot (https://www.kaggle.com/dansbecker/melbourne-housing-snapshot)
Mobile Price Classification (https://www.kaggle.com/iabhishekofficial/mobile-price-classification)

Melbourne Housing Snapshot

Описание входных данных

- Rooms – Количество комнат
- Price – Цена в долларах
- Method
  - S – Собственность была продана
  - SP – Собственность была продана с приоритетом
  - PI – Передана во владения
  - PN – Продана с приоритеом, но не раскрыта
  - SN – Продана, но не раскрыта
  - NB – Нет текущей ставки
  - VB – Цена продавца
  - W – Отозвана с аукциона (внесен залог)
  - SA – Продана с аукциона
  - SS – Продана с аукциона без раскрытия цены
  - N/A – Цена или наивысшая или неизвестна
- Type
  - br – спальни
  - h – дом, коттедж, вилла, терасса
  - u – юнит, место
  - t – таунхаус
  - dev site – правительственное здание
  - o r – другие резиденции
- SellerG – Агент недвижимости
- Date – Дата продажи
- Distance – Дистанция от аэропорта
- Regionname – Название региона

- Propertycount - Количество свойств, которые существуют в пригороде
- Bedroom2 – Царапины в спальнях (повреждения)
- Bathroom – Количество ванных комнат
- Car – Количество гаражей
- Landsize – Размер земли
- BuildingArea – Размер здания
- CouncilArea – Управляющий совет по области

Анализ данных

Типы признаков

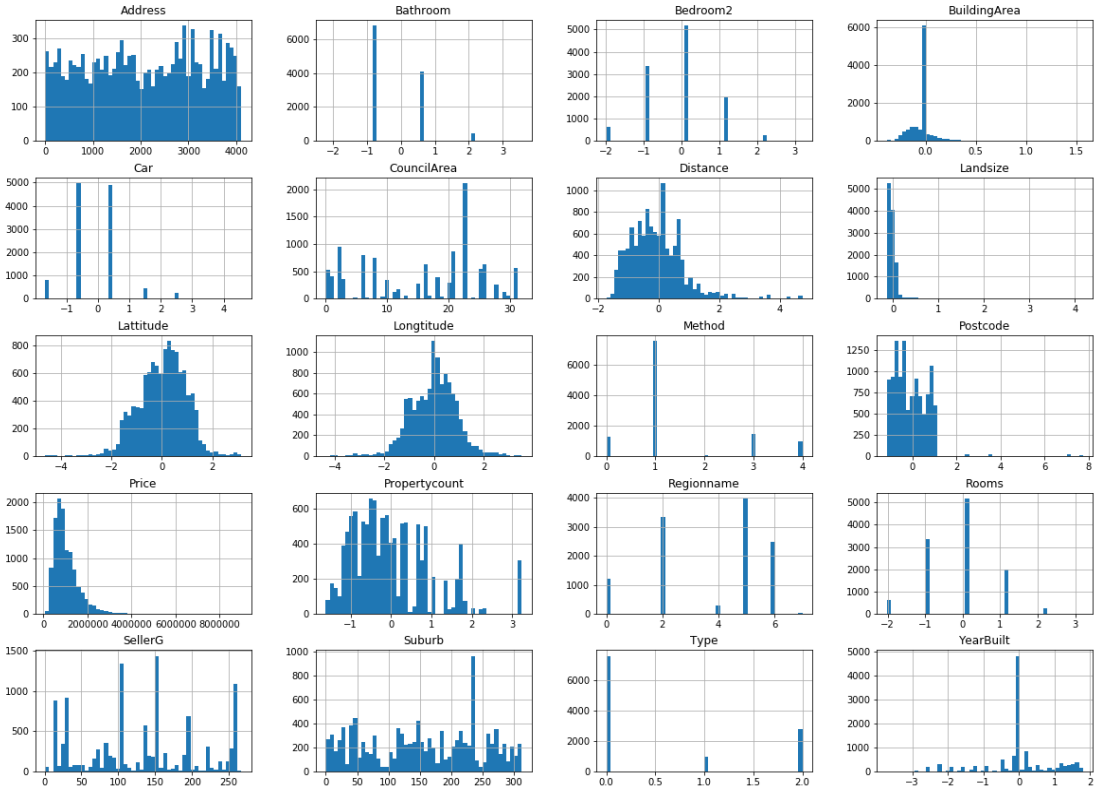
- Категориальные признаки: *Suburb, Address, Type, Method, SellerG, CouncilArea, Regionname*

- Количественные признаки: *Rooms, Distance, Postcode, Bedroom2, Bathroom, Car, Landsize, BuildingArea, YearBuilt, Lattitude, Longtitude, Propertycount*

Размер

- Строк: 13580
- столбцов: 21

Распределение призаков с числовыми полями



Решаемая задача

Предсказание признака Price.

Изначальные признаки выбранные для решения задачи

Suburb, Address, Type, Method, SellerG, CouncilArea, Regionname, Rooms, Distance, Postcode, Bedroom2, Bathroom, Car, Landsize, BuildingArea, YearBuilt, Lattitude, Longtitude, Propertycount.

Работа с категориальными признаками

Для оцифрования категориальных признаков я пользовался label encoder. Данный метод каждому из уникальных значений в текущем признаке присваивает свою метку. Также, в одном из признаков (адрес) я выделил улицу, что позволило значительно сократить количество уникальных значений.

Удаление выбросов

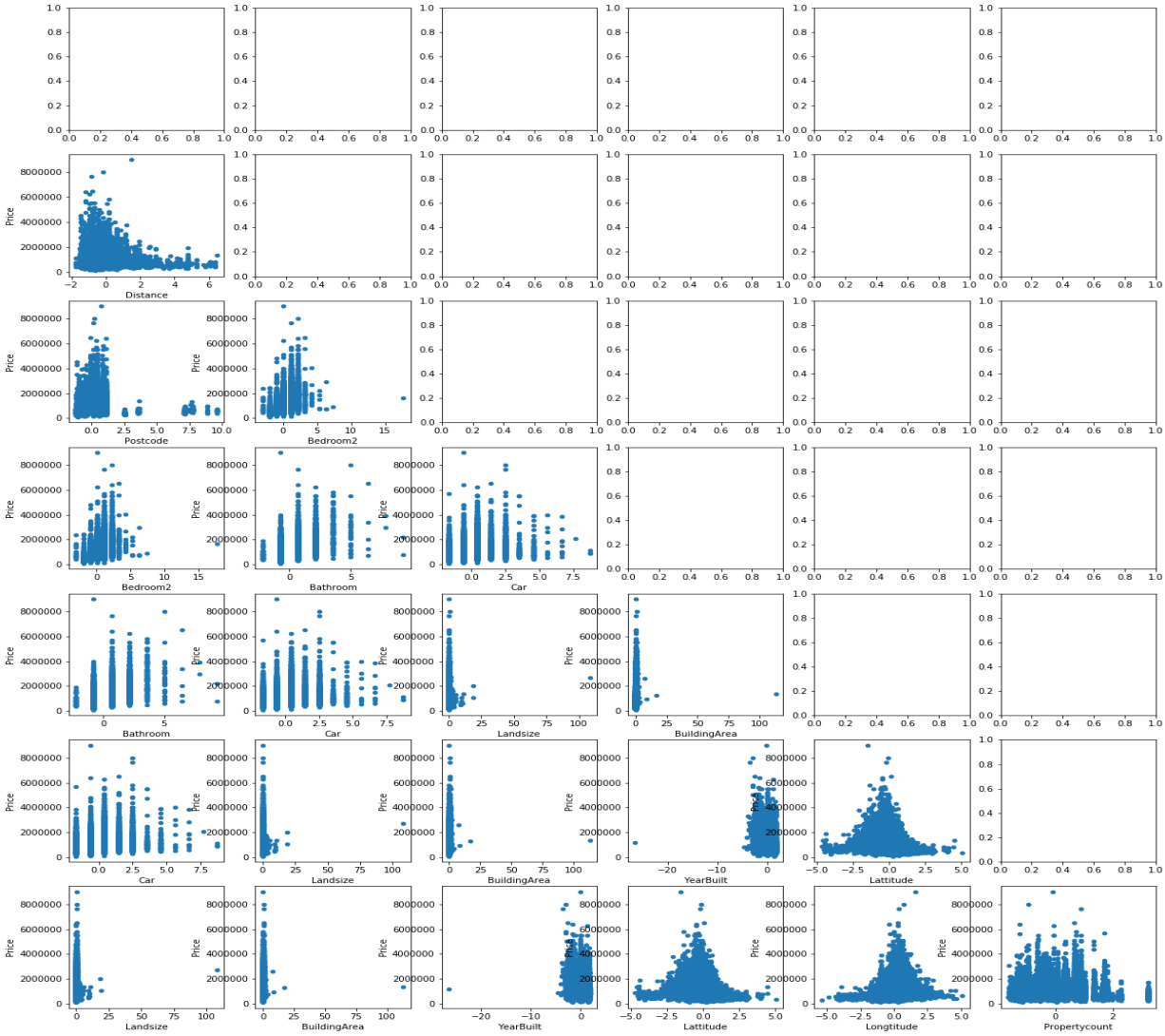
Удаление выбросов происходило при помощи алгоритма BDSCAN.

Заполнение пропусков

Пропущенные данные заполнялись на основе средних значений, а в случае категориальных признаков – как самый популярный.

Визуализация

Зависимость главного значения от всех числовых


Price
Distance
Postcode
Bedroom2
Bathroom
Car
Landsize
BuildingArea
YearBuilt
Lattitude
Longtitude
Propertycount

Mobile Price Classification

Описание входных данных

- *id – id*
- *battery_power* – Емкость батареи в МАЧ
- blue – Наличие bluetooth
- clock_speed - скорость, с которой микропроцессор выполняет инструкции
- dual_sim – Поддержка 2 симкарт
- fc – Разрещение фронтальной камеры в мегапикселях
- four_g – Наличие 4g
- int_memory – Емкость памяти в гб
- m_dep – Глубина в см
- mobile_wt – Вес в гр
- n_cores – Количество ядер
- pc – Разрешение главной камеры в мегапикселях
- ram – ОЗУ в мб
- sc_h – Высота экрана в см
- sc_w – Ширина экрана в см
- talk_time – Время в режиме разговора
- three_g – Наличие 3g
- touch_screen – Наличие тачскрина
- wifi – Наличие wifi
- price_range – К какой ценовой категории относится (0, 1, 2 ,3)

Анализ данных

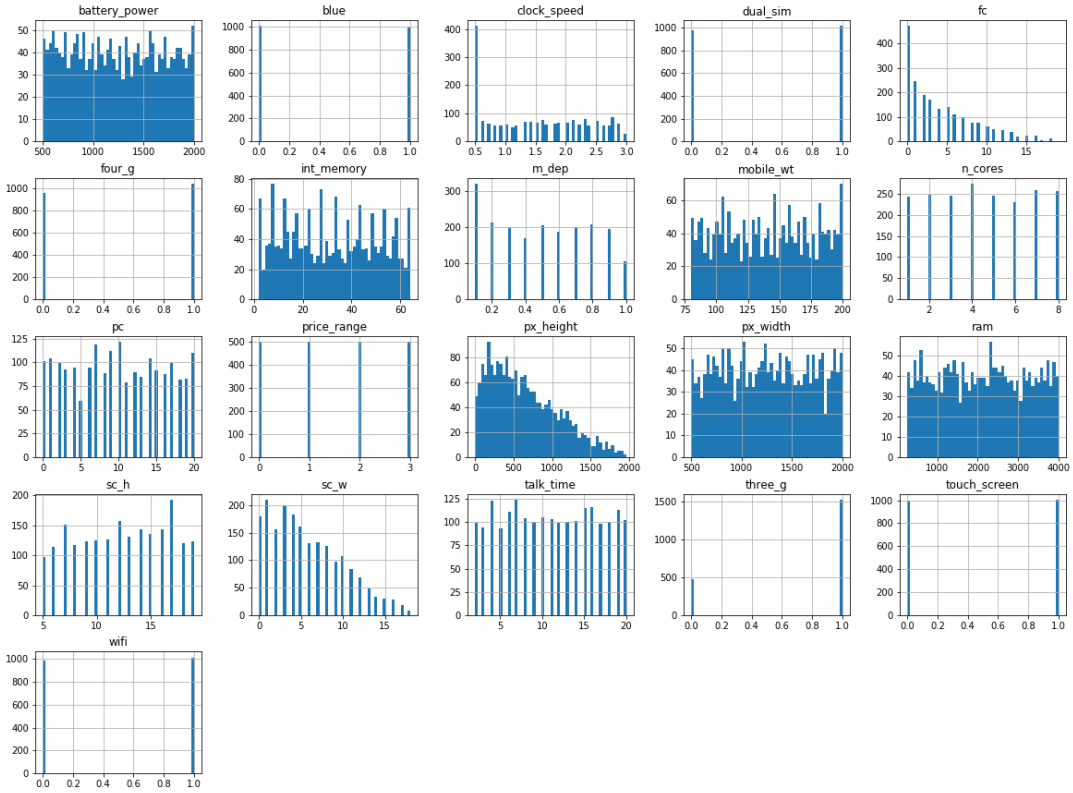
Типы признаков

- Все признаки количественные
- Бинарные: *blue, dual_sim, four_g, three_g, touch_screen, wifi*
- Иследуемое значение: *price_range*

Размер

- Строк: 2000
- Столбцов: 21

Распределение призаков с числовыми полями



Решаемая задача

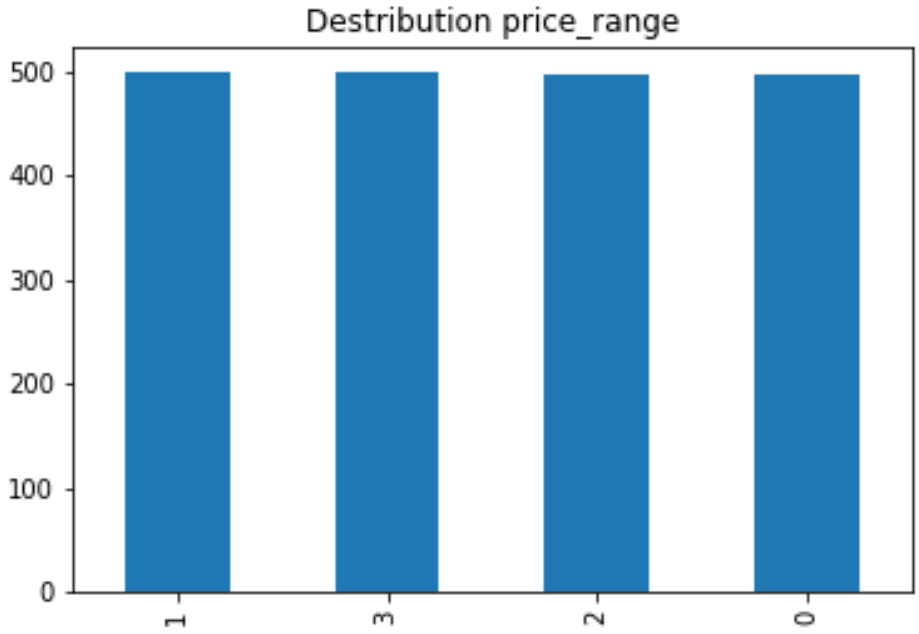
Классифицировать price_range.

Проблемы

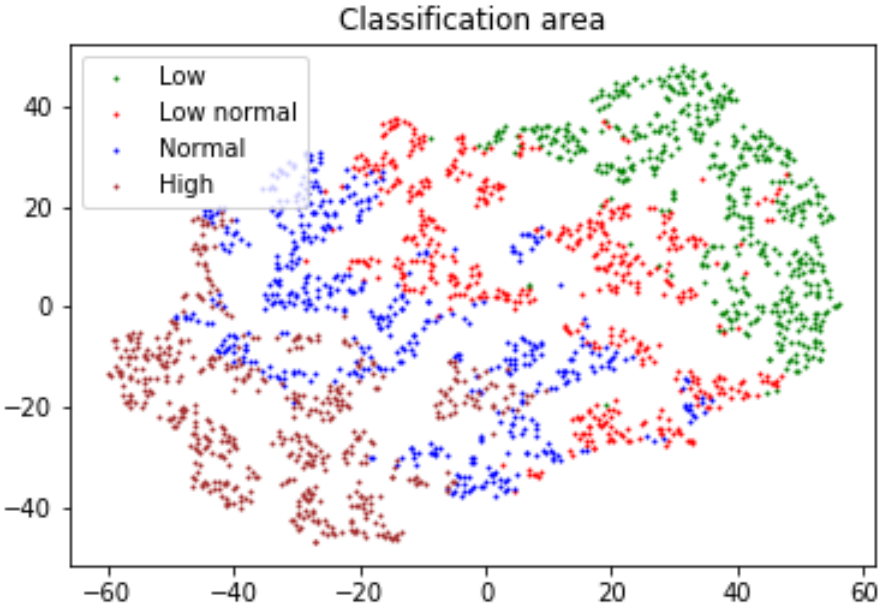
В анализе данных проблем не возникло.

Визуализация

Распределение по кластерам.



С помощью алгортима TSNE было визуализированно распределение по 4 кластерам.



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.