Московский авиационный институт

(национальный исследовательский университет)

Институт №8 «Информационные технологии и прикладная математика»

Лабораторная работа №0 по искусственному интеллекту

6 семестр

Студент: Кареткин Д.В.

Группа:М8О-301Б

Дата:

Москва, 2022

Оглавление

Постановка задачи

Найти датасет, провести первичный и визуальный анализ данных. По анализу данных сделать выводы.

Первичный анализ датасета

Был выбран датасет с информацией о характеристиках телефонов и о группах цен. Целью является классификация телефонов по ценовым группам согласно признакам телефонов.

По смыслу переменные ознают следующее:

- battery_power - заряд телефона
- blue - наличие bluetooth
- clock_speed - скорость, с которой микропроцессор выполняет инструкции
- dual_sim - возможность 2-симок
- fc - Мегапиксели фронтальной камеры
- four_g - Есть 4G или нет
- int_memory - Внутренняя память в гигабайтах
- m_dep - Мобильная глубина в см
- mobile_wt - Вес мобильного телефона
- n_cores - кол ядер
- pc - мегапиксели основной камеры
- px_height - Пиксель Разрешение Высота
- px_width - Ширина разрешения пикселей
- ram - Оперативная память в мегабайтах
- sc_h - Высота экрана мобильного телефона в см
- sc_w - Ширина экрана мобильного телефона в см
- talk_time - максимальное время работы от одного заряда батареи
- three_g - Есть 3G или нет
- touch_screen - Есть сенсорный экран или нет
- wifi - наличие wifi

Выведем данные о датасете

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   battery_power  2000 non-null   int64  
 1   blue           2000 non-null   int64  
 2   clock_speed    2000 non-null   float64
 3   dual_sim       2000 non-null   int64  
 4   fc             2000 non-null   int64  
 5   four_g         2000 non-null   int64  
 6   int_memory     2000 non-null   int64  
 7   m_dep          2000 non-null   float64
 8   mobile_wt      2000 non-null   int64  
 9   n_cores        2000 non-null   int64  
 10  pc             2000 non-null   int64  
 11  px_height      2000 non-null   int64  
 12  px_width       2000 non-null   int64  
 13  ram            2000 non-null   int64  
 14  sc_h           2000 non-null   int64  
 15  sc_w           2000 non-null   int64  
 16  talk_time      2000 non-null   int64  
 17  three_g        2000 non-null   int64  
 18  touch_screen   2000 non-null   int64  
 19  wifi           2000 non-null   int64  
 20  price_range    2000 non-null   int64  
dtypes: float64(2), int64(19)
memory usage: 328.2 KB
```

Посчитаем корреляции признаков с целевой переменной

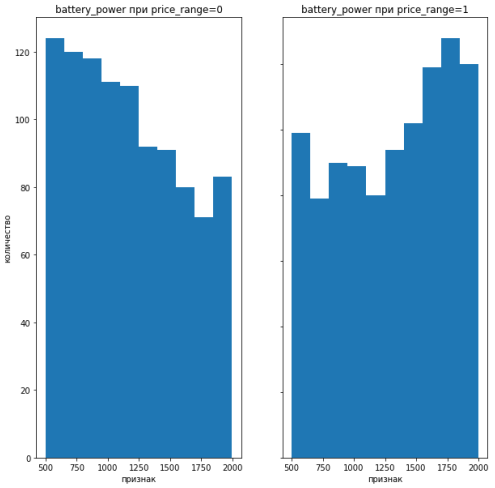
```
battery_power    0.149402
blue             0.014001
clock_speed      0.003494
dual_sim         0.009002
fc               0.022464
four_g           0.001001
int_memory       0.022132
m_dep           -0.018554
mobile_wt       -0.007968
n_cores          0.031260
pc               0.027628
px_height        0.097951
px_width         0.116703
ram              0.822354
sc_h             0.009140
sc_w             0.035359
talk_time        0.004394
three_g          0.024638
touch_screen    -0.040001
wifi             0.014001
price_range      1.000000
Name: price_range, dtype: float64
```

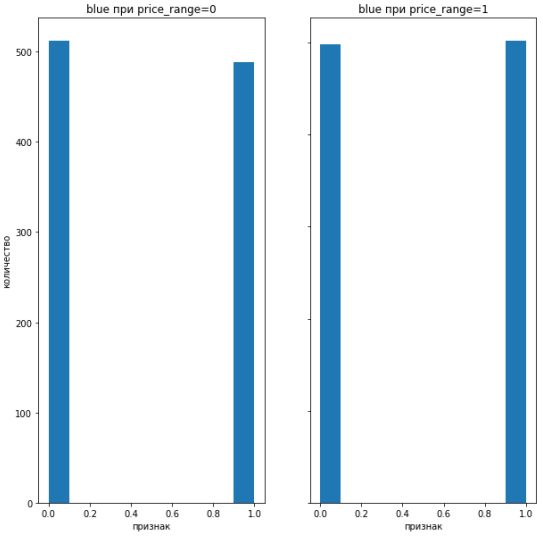
Выводы по первичному анализу

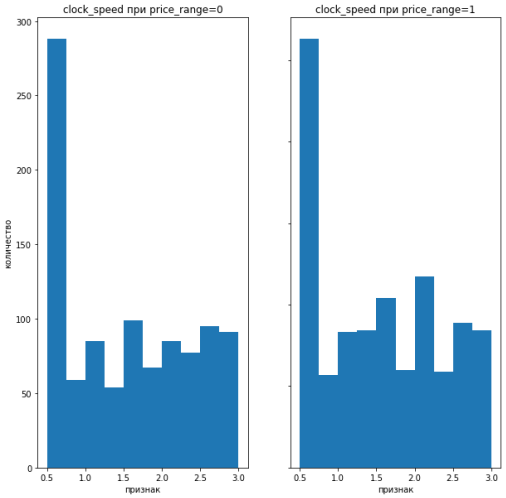
- Пропусков в данных нет
- Все данные представлены в формате int и float
- Всего 2000 объектов
- Целевая переменная сильнее всего скоррелирована с ram(коэф. корр. равенпримерно 0.82), дальше идет battery_power(0.14).

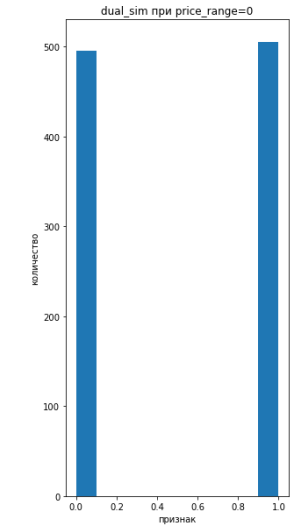
Визуализация распределений

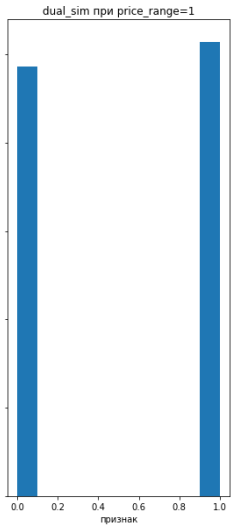
Построим гистограммы признаков в зависимости от целевой переменной.

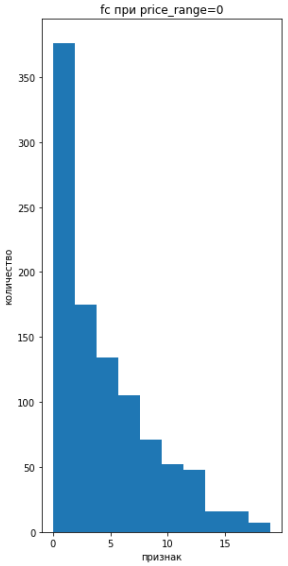


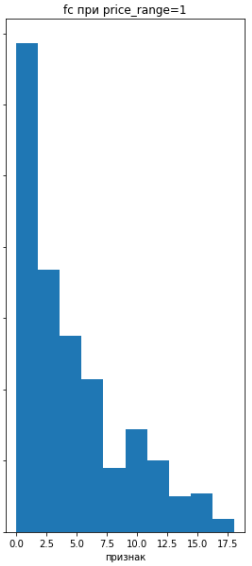


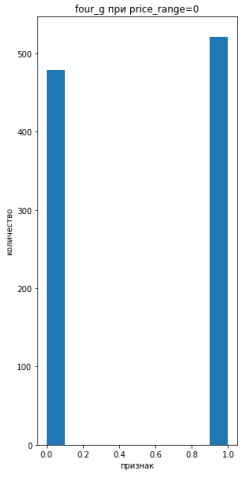


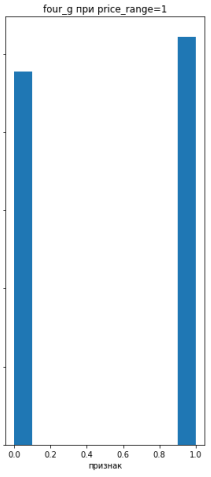

dual_sim при price_range=0
количество
признак

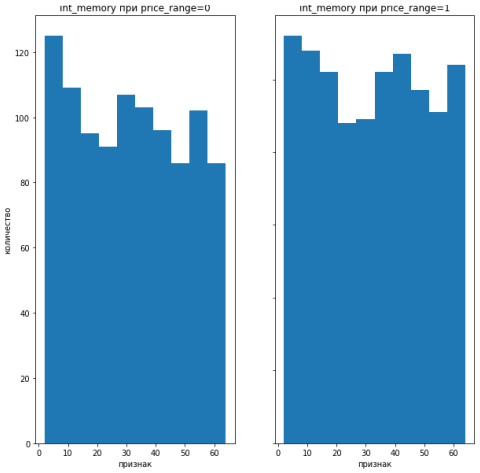

dual_sim при price_range=1
признак

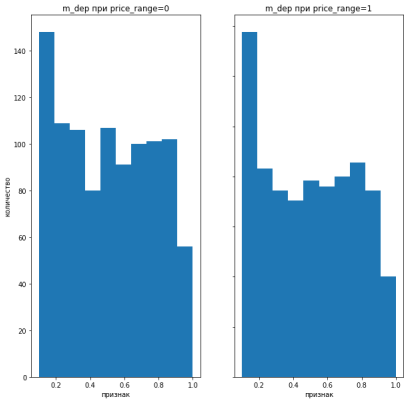

fc при price_range=0
количество
признак

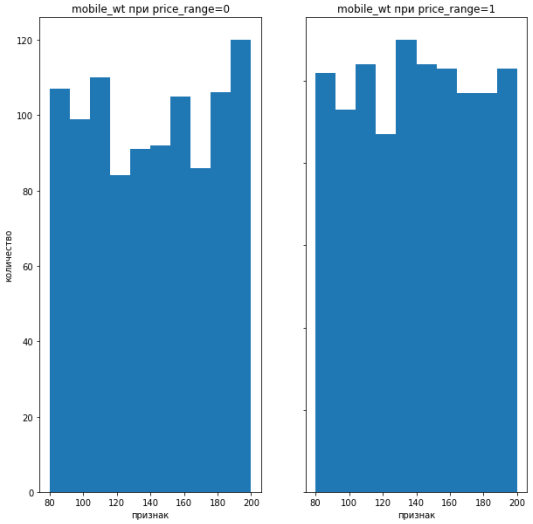

fc при price_range=1
признак

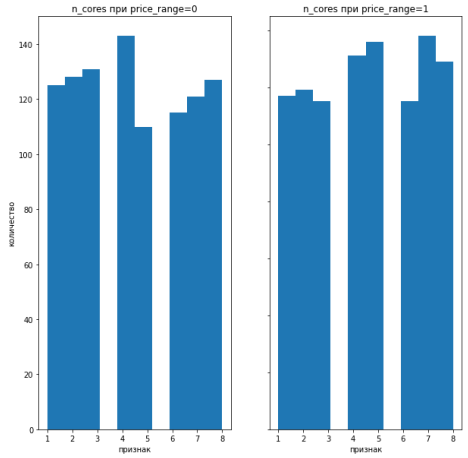

four_g при price_range=0
количество
признак

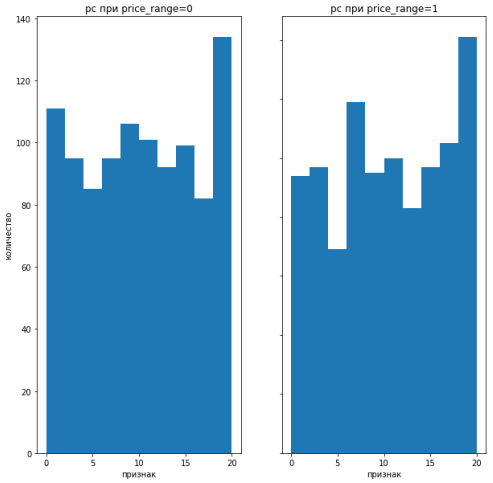

four_g при price_range=1
признак

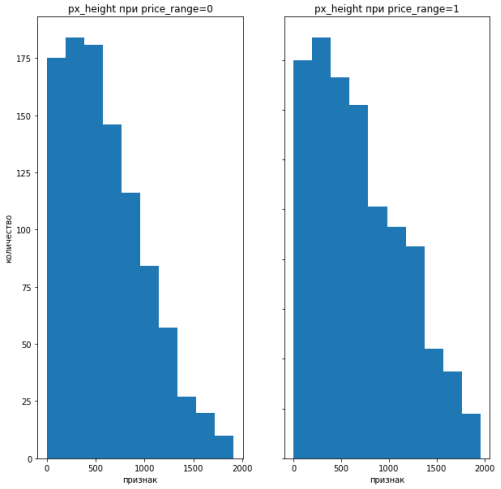

int_memory при price_range=0
int_memory при price_range=1
количество
признак

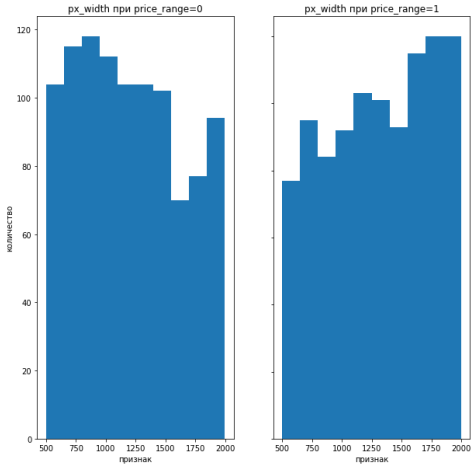

m_dep при price_range=0
m_dep при price_range=1
количество
признак

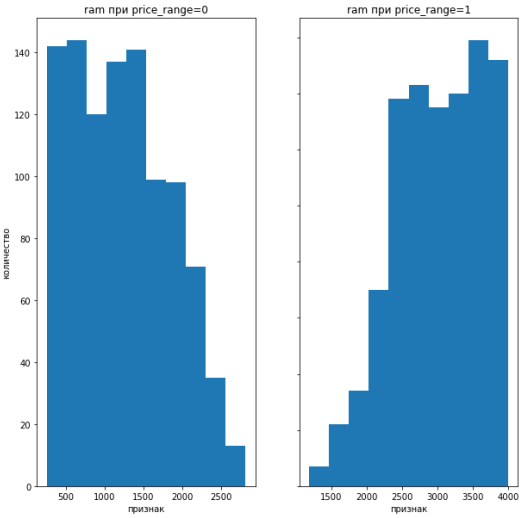

mobile_wt при price_range=0
mobile_wt при price_range=1
количество
признак

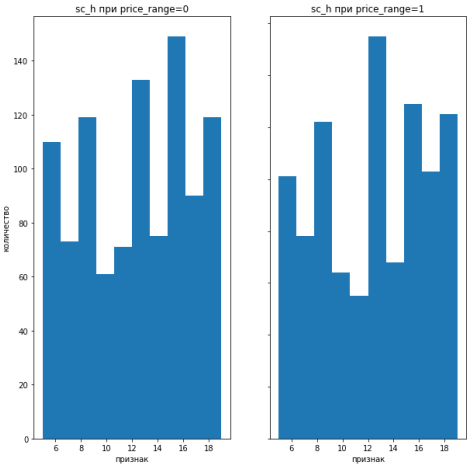

n_cores при price_range=0
n_cores при price_range=1
количество
признак
признак

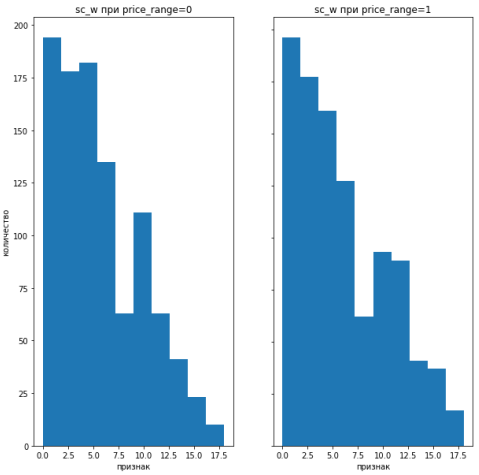

pc при price_range=0
pc при price_range=1
количество
признак
признак

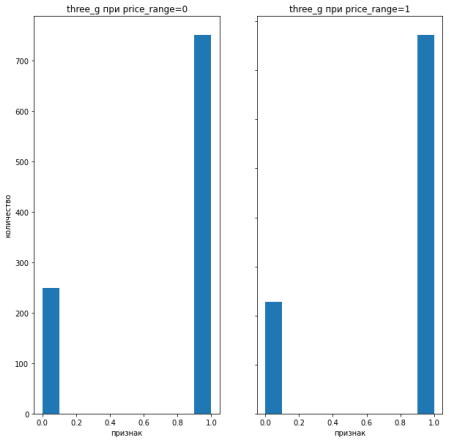

px_height при price_range=0
px_height при price_range=1
количество
признак
признак

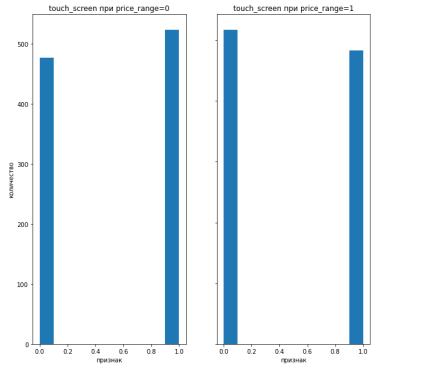

px_width при price_range=0
px_width при price_range=1
количество
признак

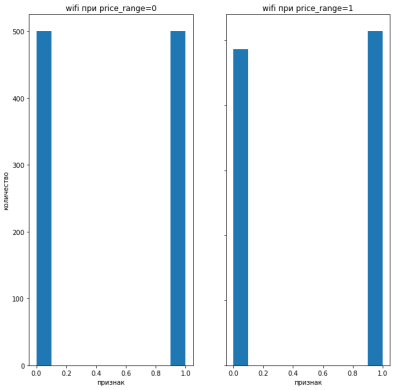

ram при price_range=0
ram при price_range=1
количество
признак


sc_h при price_range=0
sc_h при price_range=1
количество
признак









Выводы по визуальному анализу

- Признаки blue, dual_sim, four_g, int_memory, mobile_wt, touch_screen, wifi распределены примерно равномерно независимо от значения price_range
- Признак battery_power распределен зеркально при разных price_range. При этом в более дорогой группе телефонов размер батарии больше
- Согласно clock_speed в выборке представлены в основном быстрые процессоры
- Согласно n_cores количество телефон с разным числом ядер примерно одинаково в выборке

Код

```
#!/usr/bin/env python
# coding: utf-8

# # L0
# #### Импорт библиотек

# In[1]:


import pandas as pd
import matplotlib.pyplot as plt


# ### Был выбран датасет с информацией о характеристиках телефонов и о группах цен.
# ### Целью является классификация телефонов по 2 ценовым группам согласно признакам телефонов

# #### Первичный анализ данных

# In[2]:


# Загрузка данных
dataset = pd.read_csv('dataset.csv',index_col='Unnamed: 0')


# In[3]:


dataset.head()


# In[4]:


dataset.info()


# In[5]:


# Корреляция с целевой переменной
```

```
dataset.corr()['price_range']


# ### Выводы по первичному анализу
# - Пропусков в данных нет
# - Все данные представлены в формате int и float
# - Всего 2000 объектов
# - Целевая переменная сильнее всего скоррелирована с ram(коэф. корр.
равенпримерно 0.82), дальше идет battery_power(0.14).
#
# ### По смыслу переменные ознают следующее:
# - battery_power - заряд телефона
# - blue - наличие bluetooth
# - clock_speed - скорость, с которой микропроцессор выполняет инструкции
# - dual_sim - возможность 2-симок
# - fc - Мегапиксели фронтальной камеры
# - four_g - Есть 4G или нет
# - int_memory - Внутренняя память в гигабайтах
# - m_dep - Мобильная глубина в см
# - mobile_wt - Вес мобильного телефона
# - n_cores - кол ядер
# - pc - мегапиксели основной камеры
# - px_height - Пиксель Разрешение Высота
# - px_width - Ширина разрешения пикселей
# - ram - Оперативная память в мегабайтах
# - sc_h - Высота экрана мобильного телефона в см
# - sc_w - Ширина экрана мобильного телефона в см
# - talk_time - максимальное время работы от одного заряда батареи
# - three_g - Есть 3G или нет
# - touch_screen - Есть сенсорный экран или нет
# - wifi - наличие wifi

# ### Визуальный анализ данных

# In[6]:


# Функия для постраения графиков
def plot_feature(dataset,name):
    fig, axs = plt.subplots(1, 2,figsize=(10,10))
    axs[0].hist(dataset[dataset['price_range'] == 0][name])
    axs[0].set_title(f'{name} при price_range=0')
    axs[1].hist(dataset[dataset['price_range'] == 1][name])
    axs[1].set_title(f'{name} при price_range=1')


    for ax in axs.flat:
        ax.set(xlabel='признак', ylabel='количество')

    for ax in axs.flat:
        ax.label_outer()


# In[7]:


plot_feature(dataset,'battery_power')


# In[8]:


plot_feature(dataset,'blue')
```

```
# In[9]:

plot_feature(dataset,'clock_speed')

# In[10]:

plot_feature(dataset,'dual_sim')

# In[11]:

plot_feature(dataset,'fc')

# In[12]:

plot_feature(dataset,'four_g')

# In[13]:

plot_feature(dataset,'int_memory')

# In[14]:

plot_feature(dataset,'m_dep')

# In[15]:

plot_feature(dataset,'mobile_wt')

# In[16]:

plot_feature(dataset,'n_cores')

# In[17]:

plot_feature(dataset,'pc')

# In[18]:

plot_feature(dataset,'px_height')

# In[19]:
```

```
plot_feature(dataset,'px_width')


# In[20]:


plot_feature(dataset,'ram')


# In[21]:


plot_feature(dataset,'sc_h')


# In[22]:


plot_feature(dataset,'sc_w')


# In[23]:


plot_feature(dataset,'three_g')


# In[24]:


plot_feature(dataset,'touch_screen')


# In[25]:


plot_feature(dataset,'wifi')


# ### Выводы по визуальному анализу
# - Признаки blue, dual_sim, four_g, int_memory, mobile_wt, touch_screen, wifi
распределены примерно равномерно независимо от значения price_range
# - Признак battery_power распределен зеркально при разных price_range. При
этом в более дорогой группе телефонов размер батарии больше
# - Согласно clock_speed в выборке представлены в основном быстрые процессоры
# - Согласно n_cores количество телефон с разным числом ядер примерно одинаково
в выборке
#
#

# In[ ]:
```