Рубежный контроль №1 по курсу «Методы машинного обучения»

Подготовил: Студент группы ИУ5-25М Клюкин Н. А. 27.03.2024

Вариант задания

Номер варианта	Задание 1	Задание 2	Доп. требование
5	5	25	для произвольной колонки данных построить парные диаграммы (pairplot)

Описание выбранного датасета

Этот набор данных предоставляет полную информацию о поведении клиентов для типичной страницы туризма в социальных сетях. Датасет содержит в себе следующие поля:

- UserID (PK) - удаляем
- Taken_product - купил тур (да/нет)
- Yearly_avg_view_on_travel_page - Среднегодовое количество просмотров пользователем любой страницы, связанной с путешествиями
- preferred_device - Предпочтительное устройство для входа пользователя в систему
- total_likes_on_outstation_checkin_given - Общее количество лайков, поставленных пользователем при регистрации вне станции за последний год
- yearly_avg_Outstation_checkins - Среднее количество регистраций за пределами станции, выполненных пользователем
- member_in_family - Общее количество членов семьи, упомянутых пользователем в учетной записи
- preferred_location_type - Предпочтительный тип местоположения для перемещения пользователя
- Yearly_avg_comment_on_travel_page - Среднегодовые комментарии пользователя на любой странице, связанной с путешествиями
- total_likes_on_outofstation_checkin_received - Общее количество лайков, полученных пользователем при выезде за пределы станции за последний год
- week_since_last_outstation_checkin - Количество недель с момента последнего обновления пользователем регистрации вне станции
- following_company_page - Читает ли клиент страницу компании (Да или Нет)
- montly_avg_comment_on_company_page - Среднее количество комментариев пользователя на странице компании в месяц
- working_flag - работает или нет

- travelling_network_rating - Рейтинг, указывающий, есть ли у пользователя близкие друзья, которые также любят путешествовать. 1 — высокий, 4 — самый низкий
- Adult_flag - взрослый или нет
- Daily_Avg_mins_spend_on_traveling_page - Среднее время, проведенное пользователем на странице путешествия компании

Импорт библиотек

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Подгрузим датасет и продемонстрируем его содержимое
data_loaded = pd.read_csv('./data/cust_beh.csv', sep=",")
data_loaded.head()
```

```
  Taken_product  Yearly_avg_view_on_travel_page preferred_device  \
0           Yes                          307.0  iOS and Android
1            No                          367.0              iOS
2           Yes                          277.0  iOS and Android
3            No                          247.0              iOS
4            No                          202.0  iOS and Android

   total_likes_on_outstation_checkin_given
yearly_avg_Outstation_checkins  \
0                                  38570.0
1
1                                   9765.0
1
2                                  48055.0
1
3                                  48720.0
1
4                                  20685.0
1

  member_in_family preferred_location_type
Yearly_avg_comment_on_travel_page  \
0                2               Financial
94.0
1                1               Financial
61.0
2                2                   Other
92.0
3                4               Financial
56.0
4                1                 Medical
40.0

   total_likes_on_outofstation_checkin_received  \
```

```
0                                          5993
1                                          5130
2                                          2090
3                                          2909
4                                          3468

   week_since_last_outstation_checkin following_company_page  \
0                                   8                    Yes
1                                   1                     No
2                                   6                    Yes
3                                   1                    Yes
4                                   9                     No

   montly_avg_comment_on_company_page working_flag
travelling_network_rating  \
0                                  11           No
1
1                                  23          Yes
4
2                                  15           No
2
3                                  11           No
3
4                                  12           No
4

   Adult_flag  Daily_Avg_mins_spend_on_traveling_page
0         0.0                                    8.0
1         1.0                                   10.0
2         0.0                                    7.0
3         0.0                                    8.0
4         1.0                                    6.0
```

```
# Используем только некоторые признаки
cols_filter = ['Taken_product', 'preferred_device',
'Yearly_avg_view_on_travel_page', 'Yearly_avg_comment_on_travel_page',
       'travelling_network_rating']
data = data_loaded[cols_filter]
data.head()
```

```
  Taken_product preferred_device  Yearly_avg_view_on_travel_page  \
0           Yes  iOS and Android                           307.0
1            No              iOS                           367.0
2           Yes  iOS and Android                           277.0
3            No              iOS                           247.0
4            No  iOS and Android                           202.0

   Yearly_avg_comment_on_travel_page  travelling_network_rating
0                               94.0                         1
1                               61.0                         4
```

```
2                                    92.0                                    2
3                                    56.0                                    3
4                                    40.0                                    4
```

Задание 1. Для набора данных проведите кодирование одного (произвольного) категориального признака с использованием метода "one-hot encoding".

Информация из лекции

- One-hot encoding предполагает, что значение категории заменяется на отдельную колонку, которая содержит бинарные значения.
- Преимущества:
  - Простота реализации.
  - Подходит для любых моделей, так как НЕ создает фиктивное отношение порядка между значениями.
- Недостатки:
  - Расширяется признаковое пространство.

Решение

```
pd.get_dummies(data[['preferred_device']]).head()
```

```
   preferred_device_ANDROID  preferred_device_Android  \
0                     False                     False
1                     False                     False
2                     False                     False
3                     False                     False
4                     False                     False

   preferred_device_Android OS  preferred_device_Laptop  \
0                        False                    False
1                        False                    False
2                        False                    False
3                        False                    False
4                        False                    False

   preferred_device_Mobile  preferred_device_Other
preferred_device_Others  \
0                    False                   False
False
1                    False                   False
False
```

```
2                      False                     False
False
3                      False                     False
False
4                      False                     False
False

   preferred_device_Tab  preferred_device_iOS  \
0                 False                 False
1                 False                  True
2                 False                 False
3                 False                  True
4                 False                 False

   preferred_device_iOS and Android
0                              True
1                             False
2                              True
3                             False
4                              True

# Добавление отдельной колонки, признака пустых значений
pd.get_dummies(data[['preferred_device']], dummy_na=True).head()

   preferred_device_ANDROID  preferred_device_Android  \
0                     False                     False
1                     False                     False
2                     False                     False
3                     False                     False
4                     False                     False

   preferred_device_Android OS  preferred_device_Laptop  \
0                        False                    False
1                        False                    False
2                        False                    False
3                        False                    False
4                        False                    False

   preferred_device_Mobile  preferred_device_Other
preferred_device_Others  \
0                     False                     False
False
1                     False                     False
False
2                     False                     False
False
3                     False                     False
False
4                     False                     False
False
```

```
   preferred_device_Tab  preferred_device_iOS  \
0                 False                 False
1                 False                  True
2                 False                 False
3                 False                  True
4                 False                 False

   preferred_device_iOS and Android  preferred_device_nan
0                              True                 False
1                             False                 False
2                              True                 False
3                             False                 False
4                              True                 False
```

Задание 2. Для набора данных для одного (произвольного) числового признака проведите обнаружение и удаление выбросов на основе межквартильного размаха.

Информация из лекции

Межквартильный размах IQR (interquartile range, IQR) - это разность третьего квартиля и первого квартиля:

Решение

Обнаруживаем выбросы

```
def remove_outliers_iqr(data, column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    filtered_data = data[(data[column] >= lower_bound) & (data[column]
<= upper_bound)]
    return filtered_data
```

Удаление выбросов

```
data.shape
```

```
(11770, 5)
```

```
filtered_dataset = remove_outliers_iqr(data,
'Yearly_avg_view_on_travel_page')
filtered_dataset.shape
```

```
(11168, 5)
```

Построение графика по варианту

```
import seaborn as sns

sns.pairplot(filtered_dataset, vars=['Yearly_avg_view_on_travel_page',
'Yearly_avg_comment_on_travel_page'])
```

```
<seaborn.axisgrid.PairGrid at 0x239e4882f00>
```