Московский государственный технический университет им. Н.Э. Баумана.

Факультет «Информатика и управление»

Кафедра «Системы обработки информации и управления»

Отчет по лабораторной работе №3

««Подготовка обучающей и тестовой выборки, кросс-валидация и подбор гиперпараметров на примере метода ближайших соседей.»

«Курса «Технологии машинного обучения»»

Выполнила: студентка группы ИУ5-64 Светашева Ю.В

Подпись и дата:

Проверил: преподаватель каф. ИУ5 Гапанюк Ю.Е. Подпись и дата:

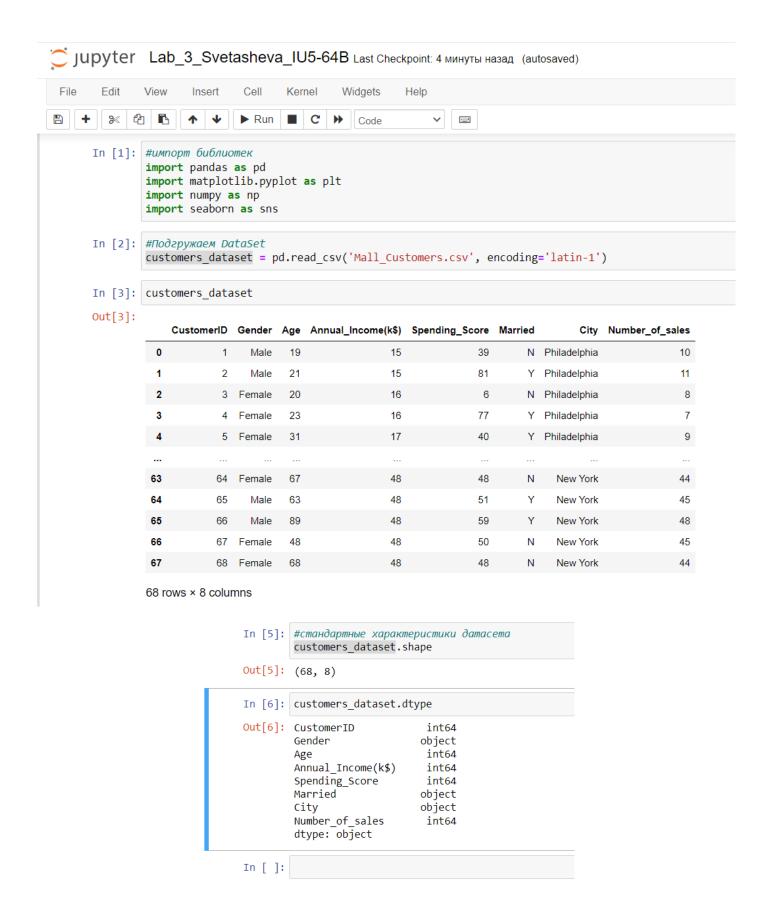
Цель лабораторной работы

Изучение способов подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.

Описание задания

- 1. Выберите набор данных (датасет) для решения задачи классификации или регрессии.
- 2. С использованием метода train_test_split разделите выборку на обучающую и тестовую.
- 3. Обучите модель ближайших соседей для произвольно заданного гиперпараметра К. Оцените качество модели с помощью подходящих для задачи метрик.
- 4. Произведите подбор гиперпараметра К с использованием GridSearchCV и/или RandomizedSearchCV и кросс-валидации, оцените качество оптимальной модели. Желательно использование нескольких стратегий кросс-валидации.
 - 5. Сравните метрики качества исходной и оптимальной моделей.

Текст программы и экранные формы



```
In [9]: #Очистка нулевых строк
             customers_dataset.isnull().sum()
    Out[9]: CustomerID
             Gender
                                  0
                                  0
             Age
             Annual_Income(k$)
             Spending_Score
                                   0
             Married
                                   0
             City
                                   0
             Number_of_sales
                                   0
             dtype: int64
    In [ ]: #Строк с нулевыми значениями не найдено
   JUDyter Lab_3_Svetasheva_IU5-64B Last Checkpoint: 26 минут назад (unsaved changes)
 File
                                                                                                                                              Pyth
                                                                                                                                  Trusted
        Edit
                View
                        Insert
                                  Cell
                                         Kernel
                                                   Widgets
                                                              Help
+
          ×
              4
                  B
                        1
                            Ψ
                                 ► Run
                                         C
                                                 *
                                                     Code
                                                                       7
      In [15]: #Кодирование категориальных признаков
                customers dataset["Gender"] = customers dataset["Gender"].astype('category')
                customers_dataset["Married"] = customers_dataset["Married"].astype('category')
                customers_dataset["City"] = customers_dataset["City"].astype('category')
                #Назначаем закодированный признакак новым столбцам с помощью accessor customers_dataset["City_cat"] = customers_dataset["City"].cat.codes
                customers_dataset["Married_cat"] = customers_dataset["Married"].cat.codes
                customers_dataset["Gender_cat"] = customers_dataset["Gender"].cat.codes
                customers_dataset
      Out[15]:
                     CustomerID Gender Age Annual_Income(k$) Spending_Score Married
                                                                                            City
                                                                                                 Number_of_sales City_cat Married_cat Gender_cat
                  0
                                   Male
                                                                                    N Philadelphia
                  1
                              2
                                   Male
                                          21
                                                            15
                                                                           81
                                                                                    Y Philadelphia
                                                                                                               11
                                                                                                                        3
                  2
                              3 Female
                                          20
                                                            16
                                                                            6
                                                                                    N Philadelphia
                                                                                                               8
                                                                                                                                   0
                                                                                                                                               0
                  3
                                          23
                                                            16
                                                                           77
                                                                                      Philadelphia
                                                                                                                        3
                                                                                                                                   1
                                                                                                                                               0
                                                                                                               9
                  4
                              5
                                          31
                                                            17
                                                                           40
                                                                                                                        3
                                                                                                                                               0
                                 Female
                                                                                    Y Philadelphia
                 63
                             64
                                 Female
                                          67
                                                           48
                                                                           48
                                                                                    N
                                                                                         New York
                                                                                                              44
                                                                                                                        2
                                                                                                                                   0
                                                                                                                                               0
                                                           48
                                                                           51
                                                                                                              45
                                                                                                                        2
                 64
                             65
                                   Male
                                          63
                                                                                    Υ
                                                                                         New York
                                                                                                                                   1
                                                                                                                                               1
                                                            48
                                                                                                               48
                 65
                             66
                                   Male
                                          89
                                                                                         New York
                 66
                             67
                                 Female
                                          48
                                                           48
                                                                           50
                                                                                    Ν
                                                                                         New York
                                                                                                              45
                                                                                                                        2
                                                                                                                                   0
                                                                                                                                               0
                                                                                                                        2
                                                                                                                                   0
                 67
                             68 Female
                                          68
                                                           48
                                                                           48
                                                                                    N
                                                                                         New York
                                                                                                              44
                                                                                                                                               0
                68 rows × 11 columns
In [29]: #Убираем ненужные столбцы, создаем новый датасет
            customers_dataset_cat = customers_dataset.drop(["City", 'Married', 'Gender'], axis=1, inplace = True)
            customers_dataset
Out[29]:
                 CustomerID
                              Age
                                     Annual_Income(k$) Spending_Score Number_of_sales City_cat Married_cat Gender_cat
              0
                            1
                                 19
                                                      15
                                                                        39
                                                                                           10
                                                                                                      3
                                                                                                                   0
                            2
              1
                                 21
                                                      15
                                                                        81
                                                                                           11
                                                                                                      3
                                                                                                                   1
              2
                            3
                                 20
                                                      16
                                                                        6
                                                                                            8
                                                                                                      3
                                                                                                                   0
                                                                                                                                0
                                                                                            7
                                                                        77
                                                                                                      3
              3
                            4
                                 23
                                                      16
                                                                                                                   1
                                                                                                                                0
                                                                        40
                                                                                            9
                                                                                                      3
                            5
                                 31
                                                      17
                                                                                                                                0
             63
                           64
                                 67
                                                     48
                                                                        48
                                                                                           44
                                                                                                      2
                                                                                                                   0
                                                                                                                                0
                                                                                                      2
                           65
                                 63
                                                     48
                                                                        51
                                                                                           45
                                                                                                                   1
             64
                                                                                                                                1
                           66
                                 89
                                                     48
                                                                        59
                                                                                           48
                                                                                                      2
                                                                                                      2
                                                                                                                   0
             66
                           67
                                 48
                                                     48
                                                                        50
                                                                                           45
                                                                                                                                0
             67
                           68
                                 68
                                                      48
                                                                        48
                                                                                           44
                                                                                                      2
                                                                                                                   0
                                                                                                                                0
```

68 rows × 8 columns

```
In [35]: #Разделение выборки на обучающую и тестовую
from sklearn.model_selection import train_test_split
y = customers_dataset['Number_of_sales']
X = customers_dataset.drop('Number_of_sales', axis=1)
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=3)
x_train
```

Out[35]:

	CustomerID	Age	Annual_Income(k\$)	Spending_Score	City_cat	Married_cat	Gender_cat
18	19	52	23	29	4	1	1
9	10	30	19	72	3	1	0
36	37	42	34	17	0	0	0
64	65	63	48	51	2	1	1
51	52	33	42	60	1	1	1
23	24	31	25	73	4	1	1
53	54	59	43	60	1	1	1
48	49	29	40	42	1	1	0
55	56	47	43	41	1	0	1
46	47	50	40	55	5	0	0
30	31	60	30	4	0	0	1
52	53	31	43	54	1	1	0
15	16	22	20	79	3	1	1
54	55	50	43	45	1	0	0
7	8	23	18	94	3	1	0
					_	•	•

```
In [36]: y_train
```

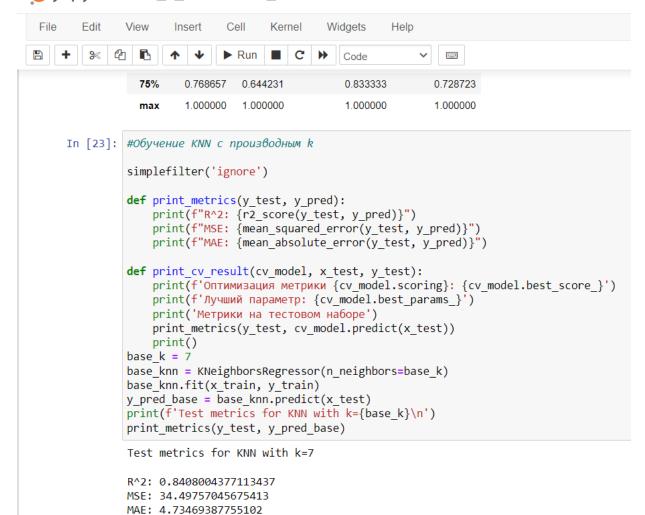
```
Out[36]: 18
               13
               10
         36
               18
         64
               45
         51
               30
         23
               15
         53
               30
         48
               31
         55
               31
         46
               27
         30
               20
         52
               37
         15
               8
         54
               32
         7
               10
         67
               44
         50
               36
         33
               21
         45
               31
         1
               11
         59
               43
         2
               8
         22
               20
         47
               32
         17
         26
               16
         14
               6
```

```
In [45]: #Масштбирование данных
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler().fit(x_train)
x_train = pd.DataFrame(scaler.transform(x_train), columns = x_train.columns)
x_test = pd.DataFrame(scaler.transform(x_test), columns = x_train.columns)
x_train.describe()
```

Out[45]:

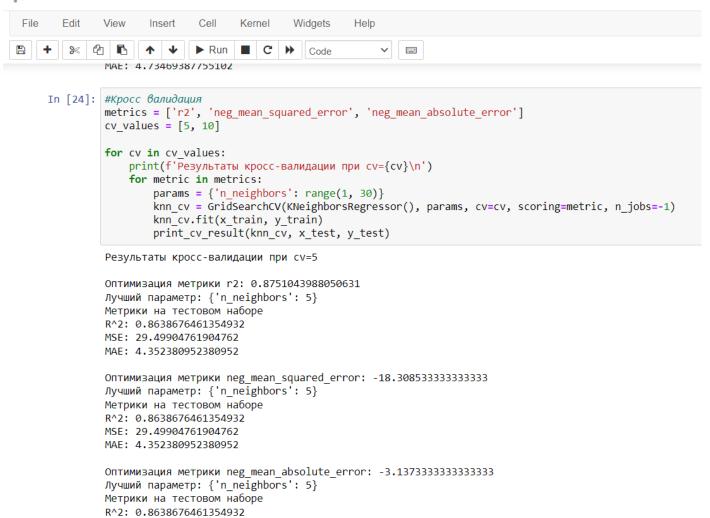
	CustomerID	Age	Annual_Income(k\$)	Spending_Score	City_cat	Married_cat	Gender_cat
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000
mean	0.510956	0.407529	0.531915	0.496831	0.493617	0.553191	0.468085
std	0.297031	0.316417	0.329406	0.276301	0.327974	0.502538	0.504375
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.276119	0.105769	0.242424	0.313830	0.200000	0.000000	0.000000
50%	0.537313	0.326923	0.575758	0.500000	0.600000	1.000000	0.000000
75%	0.768657	0.644231	0.833333	0.728723	0.800000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

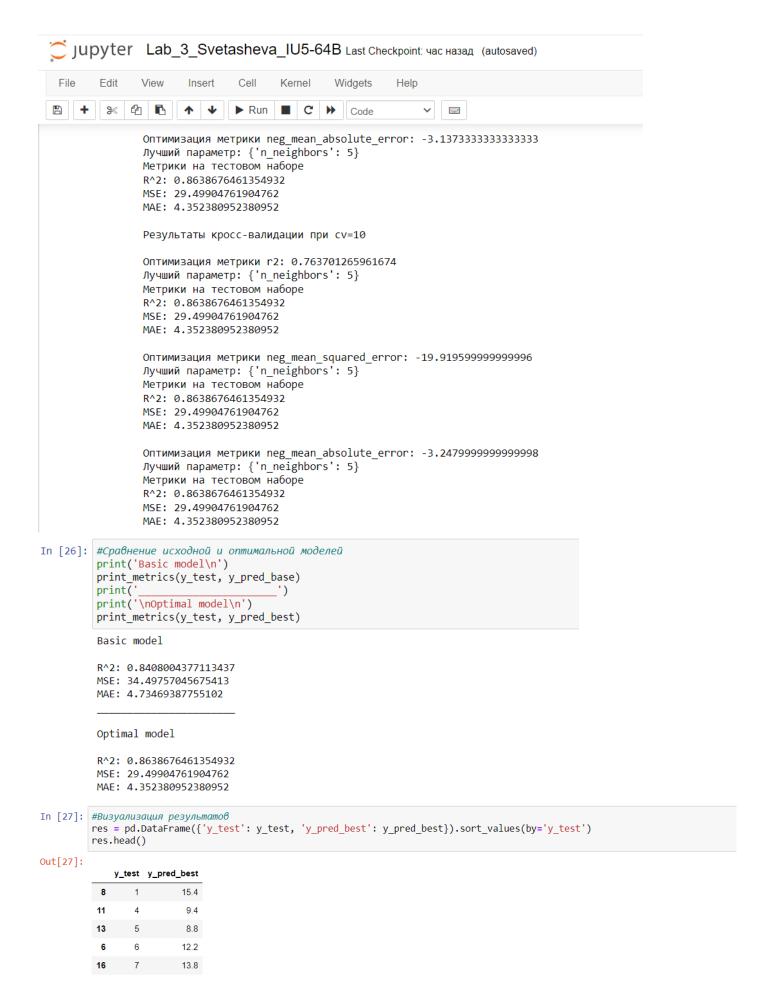
☐ Jupyter Lab_3_Svetasheva_IU5-64B Last Checkpoint: час назад (unsaved changes)



😇 Jupyter Lab_3_Svetasheva_IU5-64B Last Checkpoint: час назад (unsaved changes)

MSE: 29,49904761904762





```
In [28]: plt.figure(figsize=(16, 5))
    sns.scatterplot(range(res.shape[0]), res['y_test'], label='actual')
    sns.scatterplot(range(res.shape[0]), res['y_pred_best'], label='predicted', alpha=0.6)
    plt.ylabel('price')
    plt.xlabel('')
    plt.title(f'Best KNN model results (k={best_k})')
    plt.tick_params(axis='x', bottom=False, labelbottom=False)
    plt.show()
```

