

Московский Государственный Технический Университет
им. Н. Э. Баумана

Рулежный контроль №1
по курсу: «Технологии машинного обучения»

Разведочный анализ данных.

Выполнила:
Студентка группы ИУ5-63
Нурлыева Д.Д.

Москва
2019

Вариант №2.

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
In [15]:
data = pd.read_csv('/Users/user/Downloads/fifa.csv')
In [16]:
num_cols = []

total_count = data.shape[0]
for col in data.columns:
    # Количество пустых значений
    temp_null_count = data[data[col].isnull()].shape[0]
    dt = str(data[col].dtype)
    if temp_null_count > 0 and (dt == 'float64' or dt == 'int64'):
        num_cols.append(col)
        temp_perc = round((temp_null_count / total_count) * 100.0, 2)
        print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(col, dt, temp_null_count, temp_perc))
```

Колонка International Reputation. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Weak Foot. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Skill Moves. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Jersey Number. Тип данных float64. Количество пустых значений 60, 0.33%.

Колонка Crossing. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Finishing. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка HeadingAccuracy. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка ShortPassing. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Volleys. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Dribbling. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Curve. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка FKAccuracy. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка LongPassing. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка BallControl. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Acceleration. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка SprintSpeed. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Agility. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Reactions. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Balance. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка ShotPower. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Jumping. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Stamina. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Strength. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка LongShots. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Aggression. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Interceptions. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Positioning. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка vision. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Penalties. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Composure. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка Marking. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка StandingTackle. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка SlidingTackle. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка GKDividing. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка GKHandling. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка GKkicking. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка GKPositioning. Тип данных float64. Количество пустых значений 48, 0.26%.

Колонка GKReflexes. Тип данных float64. Количество пустых значений 48, 0.26%.

In [17]:

```
data_num = data[num_cols]
```

```
data_num
```

```
Out[17]:
```

```
18207 rows × 38 columns
```

```
In [18]:
```

```
data.head()
```

```
Out[18]:
```

```
5 rows × 89 columns
```

```
In [19]:
```

```
data.dtypes
```

```
Out[19]:
```

Unnamed: 0	int64
ID	int64
Name	object
Age	int64
Photo	object
Nationality	object
Flag	object
Overall	int64
Potential	int64
Club	object
Club Logo	object
Value	object
Wage	object
Special	int64
Preferred Foot	object
International Reputation	float64
Weak Foot	float64
Skill Moves	float64
Work Rate	object
Body Type	object
Real Face	object
Position	object
Jersey Number	float64
Joined	object
Loaned From	object
Contract Valid Until	object
Height	object
Weight	object
LS	object
ST	object
	...
Dribbling	float64
Curve	float64
FKAccuracy	float64
LongPassing	float64
BallControl	float64
Acceleration	float64
SprintSpeed	float64
Agility	float64
Reactions	float64
Balance	float64
ShotPower	float64
Jumping	float64

Stamina	float64
Strength	float64
LongShots	float64
Aggression	float64
Interceptions	float64
Positioning	float64
Vision	float64
Penalties	float64
Composure	float64
Marking	float64
StandingTackle	float64
SlidingTackle	float64
GKDividing	float64
GKHandling	float64
GKKicking	float64
GKPositioning	float64
GKReflexes	float64
Release Clause	object

Length: 89, dtype: object

In [20]:

```
data.isnull().sum()
```

Out[20]:

Unnamed: 0	0
ID	0
Name	0
Age	0
Photo	0
Nationality	0
Flag	0
Overall	0
Potential	0
Club	241
Club Logo	0
Value	0
Wage	0
Special	0
Preferred Foot	48
International Reputation	48
Weak Foot	48
Skill Moves	48
Work Rate	48
Body Type	48
Real Face	48
Position	60
Jersey Number	60
Joined	1553
Loaned From	16943
Contract Valid Until	289
Height	48
Weight	48
LS	2085
ST	2085
...	
Dribbling	48
Curve	48
FKAccuracy	48

LongPassing	48
BallControl	48
Acceleration	48
SprintSpeed	48
Agility	48
Reactions	48
Balance	48
ShotPower	48
Jumping	48
Stamina	48
Strength	48
LongShots	48
Aggression	48
Interceptions	48
Positioning	48
Vision	48
Penalties	48
Composure	48
Marking	48
StandingTackle	48
SlidingTackle	48
GKDividing	48
GKHandling	48
GKKicking	48
GKPositioning	48
GKReflexes	48
Release Clause	1564

Length: 89, dtype: int64

In [21]:

#Обработка числовых признаков

```
data[data['International Reputation'].isnull()]
```

Out[21]:

48 rows × 89 columns

In [22]:

#индексы строк с пустыми значениями

```
flt_index = data[data['International Reputation'].isnull()].index
flt_index
```

Out[22]:

```
Int64Index([13236, 13237, 13238, 13239, 13240, 13241, 13242, 13243,
13244,
           13245, 13246, 13247, 13248, 13249, 13250, 13251, 13252,
13253,
           13254, 13255, 13256, 13257, 13258, 13259, 13260, 13261,
13262,
           13263, 13264, 13265, 13266, 13267, 13268, 13269, 13270,
13271,
           13272, 13273, 13274, 13275, 13276, 13277, 13278, 13279,
13280,
           13281, 13282, 13283],
          dtype='int64')
```

In [23]:

Проверка

```
data[data.index.isin(flt_index)]
```

Out[23]:

48 rows × 89 columns

In [24]:

```
data_num[data_num.index.isin(flt_index)][ 'International Reputation' ]
```

Out[24]:

13236	NaN
13237	NaN
13238	NaN
13239	NaN
13240	NaN
13241	NaN
13242	NaN
13243	NaN
13244	NaN
13245	NaN
13246	NaN
13247	NaN
13248	NaN
13249	NaN
13250	NaN
13251	NaN
13252	NaN
13253	NaN
13254	NaN
13255	NaN
13256	NaN
13257	NaN
13258	NaN
13259	NaN
13260	NaN
13261	NaN
13262	NaN
13263	NaN
13264	NaN
13265	NaN
13266	NaN
13267	NaN
13268	NaN
13269	NaN
13270	NaN
13271	NaN
13272	NaN
13273	NaN
13274	NaN
13275	NaN
13276	NaN
13277	NaN
13278	NaN
13279	NaN
13280	NaN
13281	NaN
13282	NaN
13283	NaN

Name: International Reputation, dtype: float64

In [25]:

```
data_num_IntRep = data_num[ 'International Reputation' ]
```

```
data_num_IntRep.head()
```

```
Out[25]:
```

	International Reputation
0	5.0
1	5.0
2	5.0
3	4.0
4	4.0

```
In [26]:
```

```
from sklearn.impute import SimpleImputer
```

```
from sklearn.impute import MissingIndicator
```

```
In [27]:
```

```
indicator = MissingIndicator()
```

```
mask_missing_values_only = indicator.fit_transform(data_num_IntRep)
```

```
mask_missing_values_only
```

```
Out[27]:
```

```
array([[False],  
       [False],  
       [False],  
       ...,  
       [False],  
       [False],  
       [False]])
```

```
In [28]:
```

```
strategies=['mean', 'median', 'most_frequent']
```

```
In [29]:
```

```
impl = SimpleImputer(missing_values=np.nan, strategy='mean')
```

```
data['International Reputation'] =impl.fit_transform(data_num_IntRep)
```

```
data.isnull().sum()
```

```
Out[29]:
```

```
Unnamed: 0          0  
ID                 0  
Name              0  
Age              0  
Photo            0  
Nationality      0  
Flag            0  
Overall         0  
Potential       0  
Club           241  
Club Logo       0  
Value          0  
Wage           0  
Special        0  
Preferred Foot  48  
International Reputation  0  
Weak Foot      48  
Skill Moves    48  
Work Rate      48  
Body Type      48  
Real Face     48  
Position      60
```


Jersey Number	60
Joined	1553
Loaned From	16943
Contract Valid Until	289
Height	48
Weight	48
LS	2085
ST	2085

...

Dribbling	48
Curve	48
FKAccuracy	48
LongPassing	48
BallControl	48
Acceleration	48
SprintSpeed	48
Agility	48
Reactions	48
Balance	48
ShotPower	48
Jumping	48
Stamina	48
Strength	48
LongShots	48
Aggression	48
Interceptions	48
Positioning	48
Vision	48
Penalties	48
Composure	48
Marking	48
StandingTackle	48
SlidingTackle	48
GKDivining	48
GKHandling	48
GKKicking	48
GKPositioning	48
GKReflexes	48
Release Clause	1564

Length: 89, dtype: int64

In [30]:

#Обработка категориальных признаков

cat_cols = []

for col in data.columns:

 # Количество пустых значений

 temp_null_count = data[data[col].isnull()].shape[0]

 dt = str(data[col].dtype)

 if temp_null_count>0 and (dt=='object'):

 cat_cols.append(col)

 temp_perc = round((temp_null_count / total_count) * 100.0, 2)

 print('Колонка {}. Тип данных {}. Количество пустых значений {},

{}%.'.format(col, dt, temp_null_count, temp_perc))

Колонка club. Тип данных object. Количество пустых значений 241, 1.32%.

Колонка Preferred Foot. Тип данных object. Количество пустых значений 48, 0.26%.

Колонка Work Rate. Тип данных object. Количество пустых значений 48, 0.26%.

Колонка Body Type. Тип данных object. Количество пустых значений 48, 0.26%.

Колонка Real Face. Тип данных object. Количество пустых значений 48, 0.26%.

Колонка Position. Тип данных object. Количество пустых значений 60, 0.33%.

Колонка Joined. Тип данных object. Количество пустых значений 1553, 8.53%.

Колонка Loaned From. Тип данных object. Количество пустых значений 16943, 93.06%.

Колонка Contract valid Until. Тип данных object. Количество пустых значений 289, 1.59%.

Колонка Height. Тип данных object. Количество пустых значений 48, 0.26%.

Колонка Weight. Тип данных object. Количество пустых значений 48, 0.26%.

Колонка LS. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка ST. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RS. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LW. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LF. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка CF. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RF. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RW. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LAM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка SAM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RAM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LCM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка CM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RCM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LWB. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LDM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка CDM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RDM. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RWB. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LB. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка LCB. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка CB. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RCB. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка RB. Тип данных object. Количество пустых значений 2085, 11.45%.

Колонка Release Clause. Тип данных object. Количество пустых значений 1564, 8.59%.

In [31]:

```
cat_temp_data = data[['Work Rate']]  
cat_temp_data.head()
```

Out[31]:

	Work Rate
0	Medium/ Medium
1	High/ Low
2	High/ Medium
3	Medium/ Medium
4	High/ High

In [32]:

```
cat_temp_data['Work Rate'].unique()
```

Out[32]:

```
array(['Medium/ Medium', 'High/ Low', 'High/ Medium', 'High/ High',
      'Medium/ High', 'Medium/ Low', 'Low/ High', 'Low/ Medium',
      'Low/ Low', nan], dtype=object)
```

In [33]:

```
cat_temp_data[cat_temp_data['Work Rate'].isnull()].shape
```

Out[33]:

```
(48, 1)
```

In [34]:

```
imp2 = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
```

```
data_imp2 = imp2.fit_transform(cat_temp_data)
```

```
data_imp2
```

Out[34]:

```
array([[ 'Medium/ Medium'],
      [ 'High/ Low'],
      [ 'High/ Medium'],
      ...,
      [ 'Medium/ Medium'],
      [ 'Medium/ Medium'],
      [ 'Medium/ Medium']], dtype=object)
```

In [35]:

```
np.unique(data_imp2)
```

Out[35]:

```
array(['High/ High', 'High/ Low', 'High/ Medium', 'Low/ High', 'Low/
Low',
      'Low/ Medium', 'Medium/ High', 'Medium/ Low', 'Medium/ Medium'],
      dtype=object)
```

In [36]:

```
data['Work Rate'] = imp2.fit_transform(cat_temp_data)
```

In [37]:

```
data.isnull().sum()
```

Out[37]:

```
Unnamed: 0          0
ID                 0
Name               0
Age               0
Photo             0
Nationality       0
Flag              0
Overall           0
Potential         0
Club              241
Club Logo         0
Value             0
```

Wage	0
Special	0
Preferred Foot	48
International Reputation	0
Weak Foot	48
Skill Moves	48
Work Rate	0
Body Type	48
Real Face	48
Position	60
Jersey Number	60
Joined	1553
Loaned From	16943
Contract Valid Until	289
Height	48
Weight	48
LS	2085
ST	2085

...

Dribbling	48
Curve	48
FKAccuracy	48
LongPassing	48
BallControl	48
Acceleration	48
SprintSpeed	48
Agility	48
Reactions	48
Balance	48
ShotPower	48
Jumping	48
Stamina	48
Strength	48
LongShots	48
Aggression	48
Interceptions	48
Positioning	48
Vision	48
Penalties	48
Composure	48
Marking	48
StandingTackle	48
SlidingTackle	48
GKDivining	48
GKHandling	48
GKKicking	48
GKPositioning	48
GKReflexes	48
Release Clause	1564

Length: 89, dtype: int64

In [38]:

#Кодирование категориальных признаков

```
cat_enc = pd.DataFrame({'c1':data_imp2.T[0]})
```

cat_enc

Out[38]:

	c1
0	Medium/ Medium
1	High/ Low
2	High/ Medium
3	Medium/ Medium
4	High/ High
5	High/ Medium
6	High/ High
7	High/ Medium
8	High/ Medium
9	Medium/ Medium
10	High/ Medium
11	Medium/ Medium
12	Medium/ High
13	High/ Medium
14	Medium/ High
15	High/ Medium
16	High/ High
17	High/ High
18	Medium/ Medium
19	Medium/ Medium
20	Medium/ Medium
21	High/ High
22	Medium/ Medium
23	High/ Medium
24	Medium/ High
25	High/ Medium
26	High/ Medium
27	Medium/ High
28	Medium/ Medium
29	High/ Medium
...	...
18177	Medium/ Medium
18178	Medium/ Medium
18179	Medium/ Medium
18180	Medium/ Medium

18181	Medium/ Medium
18182	Low/ Medium
18183	Medium/ Medium
18184	Medium/ Medium
18185	Medium/ Medium
18186	Medium/ Medium
18187	High/ Medium
18188	Medium/ Medium
18189	Medium/ Medium
18190	Medium/ Medium
18191	Medium/ High
18192	Low/ Medium
18193	Medium/ Medium
18194	Medium/ Medium
18195	Medium/ Medium
18196	Medium/ Medium
18197	Medium/ Medium
18198	Medium/ Medium
18199	Medium/ High
18200	Medium/ Medium
18201	Medium/ Medium
18202	Medium/ Medium
18203	Medium/ Medium
18204	Medium/ Medium
18205	Medium/ Medium
18206	Medium/ Medium

18207 rows × 1 columns

In [39]:

#Кодирование категорий целочисленными значениями - label encoding

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

In [40]:

le = LabelEncoder()

cat_enc_le = le.fit_transform(cat_enc['c1'])

In [41]:

cat_enc['c1'].unique()

Out[41]:

array(['Medium/ Medium', 'High/ Low', 'High/ Medium', 'High/ High',
'Medium/ High', 'Medium/ Low', 'Low/ High', 'Low/ Medium',
'Low/ Low'], dtype=object)

In [42]:

np.unique(cat_enc_le)

