Демонстрационный проект

Метод классификации «Дерево решений»

Описание набора данных

- 1. X пространственная координата по оси X карты Парк Монтезиньо (Португалия) : 1 to 9
- 2. Y пространственная координата по оси Y карты Парк Монтезиньо (Португалия): 2 to 9
- 3. month месяц
- 4. day день недели
- 5. FFMC FFMC индекс оценки пожароопасности: 18.7 to 96.20
- 6. DMC DMC индекс оценки пожароопасности: 1.1 to 291.3
- 7. DC DC индекс оценки пожароопасности: 7.9 to 860.6
- 8. ISI ISI индекс оценки пожароопасности: 0.0 to 56.10
- 9. temp температура в градусах по Цельсию: 2.2 to 33.30
- 10. RH относительная влажность в %: 15.0 to 100
- 11. wind скорость ветра в км/ч: 0.40 to 9.40
- 12. rain дождь в mm/m2 : 0.0 to 6.4
- 13. area территория, пострадавшая от пожара (в Га): 0.00 to 1090.84

https://archive.ics.uci.edu/ml/datasets/Forest+Fires

Предобработка данных

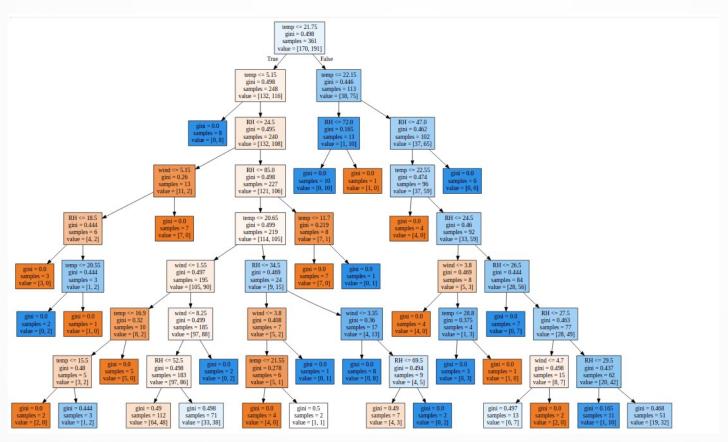
- 1. Значения месяцев заменены со строковых на целочисленные, т. к. будем использовать библиотеку Scikit-Learn.
- 2. Целевой признак из float сделаем bool.
- 3. Удалим факторы, которые не относятся к природным.
- 4. Разделим данные на целевые и входные матрицу X и вектор ответов Y.
- 5. Разделим данные на обучающие и отложенные.

```
In [422]: import pandas as pd
          from sklearn.tree import DecisionTreeClassifier
In [423]: data = pd.read_csv('forestfires.csv')
In [424]: data.head()
Out[424]:
             X Y month day FFMC DMC
                                      DC ISI temp RH wind rain area
          0 7 5
                     3 fri 86.2 26.2 94.3 5.1 8.2 51
          1 7 4
                            90.6 35.4 669.1 6.7 18.0 33
                                                       0.9 0.0 0.0
          2 7 4
                    10 sat 90.6 43.7 686.9 6.7 14.6 33
                                                      1.3 0.0 0.0
           3 8 6
                     3 fri 91.7 33.3 77.5 9.0 8.3 97
                                                       4.0 0.2 0.0
          4 8 6
                     3 sun 89.3 51.3 102.2 9.6 11.4 99 1.8 0.0 0.0
In [425]: data['burned'] = data['area'] > 0
In [426]: data.drop(['area','day', 'FFMC', 'DMC', 'DC', 'ISI', 'X', 'Y', 'month'], axis=1, inplace=True)
In [427]: data.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 517 entries, 0 to 516
          Data columns (total 5 columns):
                    517 non-null float64
          temp
                    517 non-null int64
          RH
                    517 non-null float64
          wind
                    517 non-null float64
          rain
          burned
                    517 non-null bool
          dtypes: bool(1), float64(3), int64(1)
          memory usage: 16.7 KB
In [428]: y = data['burned']
In [429]: x = data.drop('burned', axis=1)
In [430]: from sklearn.model selection import train test split, cross val score
          import numpy as np
In [431]: X train, X valid, y train, y valid = train test split(x, y,
                                                                test size=0.3,
                                                                random state=4)
```

Кросс-валидация и подбор параметров

```
In [432]: first tree = DecisionTreeClassifier(random state=4)
In [433]: np.mean(cross val score(first tree, X train, y train, cv=5))
Out[433]: 0.5403729071537291
In [434]: from sklearn.model selection import GridSearchCV
In [435]: tree params = {'max_depth': np.arange(1, 11)}
In [436]: tree grid = GridSearchCV(first tree, tree params, cv=5, n jobs=-1)
In [437]: tree grid.fit(X train, y train);
          /home/rravilov/PycharmProjects/pandastraining/venv/lib/python3.6/site-packages/sklearn/model selection/ search.py:8
          13: DeprecationWarning: The default of the 'iid' parameter will change from True to False in version 0.22 and will
          be removed in 0.24. This will change numeric results when test-set sizes are unequal.
            DeprecationWarning)
In [438]: tree grid.best score , tree grid.best params
Out[438]: (0.556786703601108, {'max depth': 8})
```

Прогноз для отложенной выборки и оценка метрикой Accuracy score



Выводы: природные факторы не влияют на лесные пожары парка Монтезиньо на севере Португалии.

