

Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

Факультет «Информатика и системы управления» Кафедра «Системы обработки информации и управления»

Рубежный контроль № 2 по курсу "Технологии машинного обучения" по теме "Технологии разведочного анализа и обработки данных" Вариант 7

Выполнил:

студент группы ИУ5-63Б

Волгина А. Д.

25.05.21

Проверил:

Гапанюк Ю.Е.

Задание

Для заданного набора данных (по Вашему варианту) постройте модели классификации или регрессии (в зависимости от конкретной задачи, рассматриваемой в наборе данных). Для построения моделей используйте методы 1 и 2 (по варианту для Вашей группы). Оцените качество моделей на основе подходящих метрик качества (не менее двух метрик). Какие метрики качества Вы использовали и почему? Какие выводы Вы можете сделать о качестве построенных моделей? Для построения моделей необходимо выполнить требуемую предобработку данных: заполнение пропусков, кодирование категориальных признаков, и т.д.

Методы:

Дерево решений и случайный лес

Набор данных:

https://www.kaggle.com/mohansacharya/graduate-admissions (файл Admission_Predict_Ver1.1.csv)

Импортируем модули для работы и загрузим набор данных в переменную data:

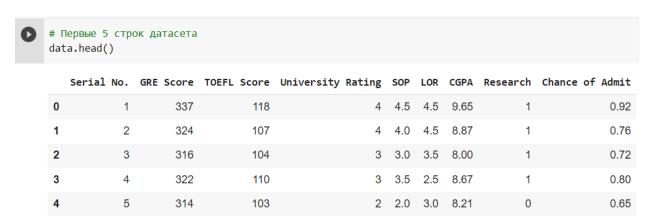
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.ensemble import RandomForestRegressor
%matplotlib inline
```

```
[2] data = pd.read_csv('Admission_Predict_Ver1.1.csv', sep=",")
```

Убедимся, что в нём нет пропусков:

```
[3] data.isnull().sum()
    Serial No.
                          0
                          0
    GRE Score
    TOEFL Score
                          0
    University Rating
                          0
    SOP
                          0
    LOR
                          0
    CGPA
    Research
                          0
    Chance of Admit
                          0
    dtype: int64
```

Взглянем на наш датасет:



Сделаем разделение на признаки и целевые значения:

```
y = data[data.columns[-1]].values
X = data[data.columns[1:-1]].values
```

Целевой признак – chance of admit, а номер студента выбрасываем из матрицы признаков, т.к. от этого параметра ничего не зависит.

Поделим выборку на тренировочную и тестовую:

```
X_train, X_test, y_train, y_test = train_test_split(
          X, y, test_size=0.2, random_state=42)
```

В данном датасете нужно по заданным признакам определить, каковы шансы на поступление у студента. Это задача регрессии, поэтому для оценки качества моделей будем использовать метрики mse, mae и r-squared.

Дерево решений:

```
mdl1 = DecisionTreeRegressor(max_depth=10)
mdl1.fit(X_train, y_train)
y1 = mdl1.predict(X_test)
print("MSE: ", mean_squared_error(y1, y_test))
print("MAE: ", mean_absolute_error(y1, y_test))
print("R-squared:", mdl1.score(X_test, y_test))
```

MSE: 0.007713640340765936 MAE: 0.0599872222222222 R-squared: 0.6228048733121793

Качество модели достаточно неплохое. Метрика R-squared показывает зависимость дисперсий зависимой переменной от независимых, а по тае и тясе видно, что в среднем отклонение предсказанных значений от настоящих невелико.

Случайный лес:

```
mdl2 = RandomForestRegressor(max_depth=10)
mdl2.fit(X_train, y_train)
y2 = mdl2.predict(X_test)
print("MSE: ", mean_squared_error(y2, y_test))
print("MAE: ", mean_absolute_error(y2, y_test))
print("R-squared:", mdl1.score(X_test, y_test))
```

MSE: 0.004401231087855343 MAE: 0.043403633058803764 R-squared: 0.6228048733121793

Аналогично вышло и в этой модели — качество хорошее. Немного лучше, чем в дереве решений, но разница незначительная.