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)»

Факультет: Информатика и системы управления

Кафедра: Теоретическая информатика и компьютерные технологии

Домашнее задание №2

«Анализ поведения клиентов компании»

по дисциплине «Моделирование»

Работу выполнил

студент группы ИУ9-82Б

Жук Дмитрий

Москва, 2023

Цель работы

Целью данной работы является построение и сравнение различных моделей в задаче прогнозирования ухода клиентов из компании.

Задание

Аналитики коммерческой компании заметили, что из нее стали уходить клиенты. В наличии есть данные о поведении клиентов и расторжении договоров с компанией за прошлые периоды. Нужно спрогнозировать, уйдёт ли конкретный клиент в ближайшее время или нет. Сравните применимость трех моделей для задачи классификации: логистическая регрессия, дерево принятия решений, случайный лес. Постройте модель с предельно большим значением F1-меры, метрика не должна превышать до 0.65. Предварительно проверьте F1-меру на тестовой выборке.

Признаки в наборе данных:

- RowNumber – индекс строки в данных
- CustomerId – уникальный идентификатор клиента
- Surname – фамилия
- Score – рейтинг клиента
- Geography – страна проживания
- Gender – пол
- Age – возраст
- Tenure – сколько лет человек является клиентом компании
- Balance – баланс, доступный для оплаты услуг компании на карте
- NumOfProducts – количество продуктов компании, используемых клиентом
- Has – наличие привилегий

- IsActiveMember – активность клиента
- EstimatedSalary – предполагаемая зарплата
- Exited – факт ухода клиента

Теория

Задача классификации заключается в разделении множества объектов на классы, где известна принадлежность некоторого подмножества объектов к определенным классам. Для остальных объектов необходимо определить их принадлежность к имеющимся классам.

Задача об уходе клиентов от компании является задачей бинарной классификации, где объекты разделяются на два класса: клиенты, покинувшие компанию, и клиенты, оставшиеся в компании. Для решения таких задач можно использовать различные модели, такие как логистическая регрессия, дерево и случайный лес. Для оценки результатов работы моделей используется метрику F1-мера, которая определяется следующим образом:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall},$$

где $recall$ – полнота (количество верно предсказанных положительных результатов, деленное на сумму верно и неверно предсказанных положительных результатов и неверно предсказанных отрицательных результатов), $precision$ – точность (количество верно предсказанных положительных результатов деленное на сумму верно и неверно предсказанных положительных результатов);

Реализация

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import f1_score, mean_squared_error
from sklearn.pipeline import Pipeline

data = pd.read_csv('dataset.csv', sep=',')
data = data.drop('Unnamed: 0', axis=1).drop('Id', axis=1)

feature_columns = data.columns.to_list()[1:-1]
target_column = data.columns.to_list()[-1]
id_column = data.columns.to_list()[0]

data.dropna(inplace=True)

data['Gender'] = LabelEncoder().fit_transform(data['Gender'])
data['Geography'] = LabelEncoder().fit_transform(data['Geography'])
data.drop(['CustomerId', 'Surname'], axis=1, inplace=True)

X = data.drop('Exited', axis=1)
y = data['Exited']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=12345)

pipelines = [
    ('Логистическая регрессия', Pipeline([
        ('scaler', StandardScaler()),
        ('classifier', LogisticRegression(solver='liblinear',
class_weight='balanced'))
    ])),
    ('Дерево принятия решений', Pipeline([
        ('scaler', StandardScaler()),
        ('classifier', DecisionTreeClassifier())
    ])),
    ('Случайный лес', Pipeline([
        ('scaler', StandardScaler()),
        ('classifier', RandomForestClassifier())
    ]))
]

index = 1
for name, pipeline in pipelines:
  pipeline.fit(X_train, y_train)
  y_pred = pipeline.predict(X_test)
  print(f"# {name}")
  print(f"  F1 score = {f1_score(y_test, y_pred)}")
  print(f"  MSE      = {mean_squared_error(y_test, y_pred)}")
  print(f"  accuracy = {pipeline.score(X_test, y_test)}")
  print()
```

Результаты

Некоторые признаки в исходном наборе данных не имеют влияния на факт ухода клиента из компании. Эти признаки включают RowNumber (индекс строки в данных), CustomerId (уникальный идентификатор клиента) и фамилию клиента (Surname). Поэтому они были удалены из набора данных. Кроме того, в наборе данных есть категориальные признаки, такие как страна проживания (Geography) и пол (Gender), которые были закодированы числами с помощью sklearn.preprocessing.LabelEncoder, так как модели машинного обучения, как правило, требуют числовых данных.

На рисунке 1 представлен фрагмент набора данных после предварительной обработки.

	Score	Geography	Gender	Age	Tenure	Balance	NumOfProducts	Has	IsActiveMember	EstimatedSalary	Exited
0	619	0	0	42	1.0	83807.86	1	0	1	112542.58	0
1	608	2	0	41	8.0	159660.80	3	1	0	113931.57	1
2	502	0	0	42	1.0	0.00	2	0	0	93826.63	0
3	699	0	0	39	2.0	125510.82	1	1	1	79084.10	0
4	850	2	0	43	8.0	113755.78	2	1	0	149756.71	1
...	...	...	...	...	...	...	...	...	...	...	...
94	800	0	0	39	5.0	0.00	2	1	0	96270.64	0
95	771	0	1	35	10.0	57369.61	1	1	1	101699.77	0
96	516	0	1	36	7.0	0.00	1	0	1	42085.58	1
97	709	0	0	42	3.0	75075.31	2	1	0	92888.52	1
99	792	0	0	39	10.0	129845.26	1	1	1	96444.88	0

Рисунок 1 – Фрагмент обработанного набора данных

В таблице 1 приведено сравнение результатов моделей.

Станок	F_1	Точность
Логистическая регрессия	0,54	0,73
Дерево принятия решений	0,5	0,79
Случайный лес	0,28	0,79

Таблица 1 – Результат модели

Вывод

В процессе выполнения данной лабораторной работы были реализованы модели классификации, включая логистическую регрессию, дерево принятия решений и случайный лес.

Многократное тестирование этих моделей показало значительный разброс в значениях F1-меры, не превышающий 10%, из-за недостаточного объема данных в тестовом наборе. Логистическая регрессия предпочтительнее для решения данной задачи, поскольку обладает высокой точностью и стабильностью. Хотя дерево принятия решений иногда показывает наибольшую F1-меру, она может быть скомпрометирована из-за редких случаев ухода клиентов из компании. Сравнение дерева принятия решений и случайного леса показывает, что их точность может быть одинаковой, но значение F1-меры может различаться, что помогает исключить ситуации, когда высокая точность достигается за счет неучтенных признаков. Ожидаемо, показатели случайного леса оказались ниже, поскольку этой модели требуется больше данных для обучения.