Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования
Московский государственный технический университет имени Н.Э. Баумана

Домашнее задание №2
«Анализ поведения клиентов компании»
по курсу «Моделирование»

Выполнил:
студент группы ИУ9-82Б
Егоров Алексей

Москва, 2023

1. Цель работы

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

2. Постановка задачи

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

3. Теоретические сведения

Задача классификации описывается следующим образом. Множество объектов предметной области некоторым образом разделены на классы. Для конечного подмножества известна принадлежность к определенным классам. На

основе этой информации необходимо отнести оставшиеся объекты к существующим классам.

Поставленная задача об уходе клиентов является задачей бинарной классификации, так как она подразумевает разделение объектов на два класса: клиент, покинувший компанию, и клиент, оставшийся в компании.

Для решения задач предлагается использовать следующие модели: логистическая регрессия, дерево, случайный лес. Для оценки результата работы моделей алгоритма, используется F1-мера, которая определяется следующим образом:

$$F1 = 2\frac{precision \cdot recall}{precision + recall},$$

где recall — полнота модели, определяемая как доля истинно положительных срабатываний к сумме истинно положительных и ложно отрицательных срабатываний; precision — точность модели, определяемая как доля истинно положительных срабатываний среди всех положительных срабатываний

4. 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 = [
    ('Logistic Regression', Pipeline([
        ('scaler', StandardScaler()),
        ('classifier', LogisticRegression(solver='liblinear',
        ↪  class_weight='balanced'))
    ])),
    ('Decision Tree', Pipeline([
        ('scaler', StandardScaler()),
        ('classifier', DecisionTreeClassifier())
    ])),
    ('Random Forest', Pipeline([
        ('scaler', StandardScaler()),
        ('classifier', RandomForestClassifier())
    ]))
]
for name, pipeline in pipelines:
  pipeline.fit(X_train, y_train)
```

```
    y_pred = pipeline.predict(X_test)
    f1 = f1_score(y_test, y_pred)
    print(f"{name}: F1 score = {f1}")
    print(f"{name}: MSE = {mean_squared_error(y_test, y_pred)}")
    print(f"{name}, accuracy = {pipeline.score(X_test, y_test)}")
```

5. Результат

Изначальный набор данных содержит признаки, которые заведомо не влияют на факт ухода клиента из компании:

— RowNumber – индекс строки в данных

— CustomerId – уникальный идентификатор клиента

— Surname – фамилия

Данные признаки были удалены из набора данных.

Также набор данных содержит некоторые категориальные признаки, которым, с помощью sklearn.preprocessing.LabelEncoder, было присвоено числовое значение, т.к. модели изначально предполагают работу именно с числовыми данными:

— Geography – страна проживания

— Gender – пол

На рисунке 1 представлен фрагмент набора данных после предварительной обработки.

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

Таблица 1 — Результаты моделей

	F1	Точность
Логистическая регрессия	0.54	0.73
Дерево принятия решений	0.6	0.79
Случайный лес	0.28	0.79

6. Вывод

В ходе данной лабораторной работы были реализованы следующие модели классификации: логистическая регрессия, дерево принятия решений, случайный лес.

Многократное повторение классификации показало достаточно большой разброс значений F1-меры (в пределах 10%). Это связано с недостаточным объемом данных в тестовом наборе. Несмотря на то, что дерево выбора иногда показывает наиболее высокое значение F1-меры, логистическа регрессия является предпочтительной для решения поставленной задачи, так как она обладает сравнительно высокими показателями и стабильностью. Сравнение дерева принятия решений и случайного леса демонстрирует отличие F1-меры от точности: при одинаковой точности, F1-меры моделей имеют различные значения. F1-мера помогает исключить ситуацию, когда высокая точность достигается за счёт того, что клиенты редко покидают компанию, и модель классификации независимо от признаков относит клиента к оставшимся в компании. Показатели случайного леса ожидаемо оказались ниже показателей дерева принятия решений, так как этой модели требуется больше данных для обучения.