

**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

**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:

где – полнота (количество верно предсказанных положительных результатов, деленное на сумму верно и неверно предсказанных положительных результатов и неверно предсказанных отрицательных результатов), – 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.

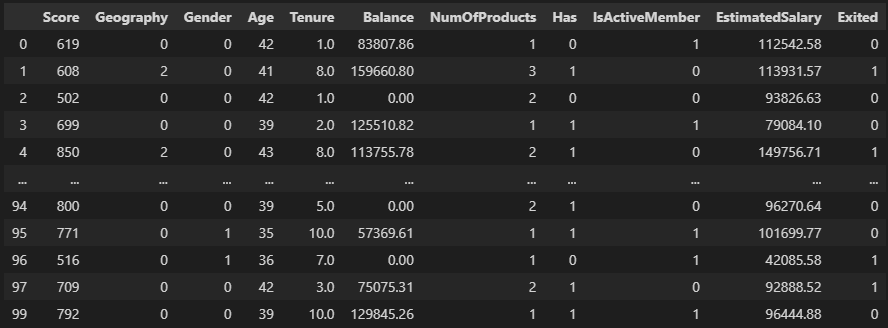


Рисунок 1 – Фрагмент обработанного набора данных

В таблице 1 приведено сравнение результатов моделей.

|  |  |  |
| --- | --- | --- |
| Станок |  | 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.