ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ассистент |  |  |  | В.В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ |
| Применение методов классификации |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4917 |  |  |  | Д.В. Потемкин |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2022

1. **Цель работы**

изучить алгоритмы и методы классификации на практике.

**Вариант 15:** 5gym\_churn

Информация о клиентах фитнес клуба:

1. Пол
2. Близкое расположение
3. Сотрудник компании партнера
4. По промо друзей
5. Указан ли телефон
6. Длительность текущего абонемента
7. Посещение групповых занятий
8. Возраст
9. Средние траты на дополнительные услуги
10. Количество месяцев до окончания абонемента
11. время с момента первого обращения в фитнес-центр (в месяцах)
12. Средняя частота посещений
13. Средняя частота посещений в текущем месяце
14. Покинул ли человек фитнес клуб
15. **Ход работы**

Импортируем датафрейм 5gym\_churn.csv и проведём предварительную обработку данных (Рисунок 1):

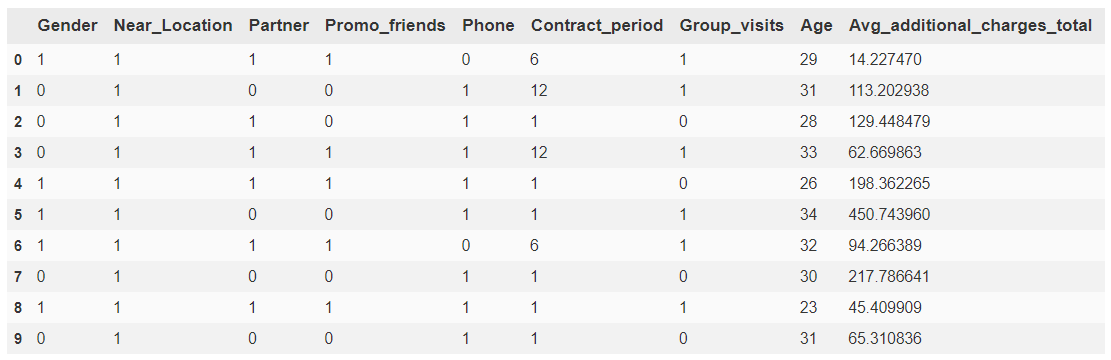


Рисунок 1 – Импортированный датафрейм

Целевым признаком выберем колонку с информацией покинул ли человек фитнес клуб или нет – Churn. Уберём из таблицы с признаками столбцы, сильно коррелирующие друг с другом, а также значения, слишком слабо коррелирующие с целевым признаком. Разобъём набор данных на тренировочной и тестовый датасеты с помощью train\_test\_split.

Разработаем предсказательную модель качественного отклика методом k-ближайших соседей и подсчитаем метрики "Accuracy", "Precision", "Recall", "Balanced accuracy", "F1 score" (Рисунок 2):

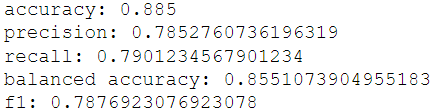


Рисунок 2 – Метрики оценки классификации методом k-ближайших соседей

Высчитаем значения матрицы ошибок для метода k-ближайших соседей (Рисунок 3):

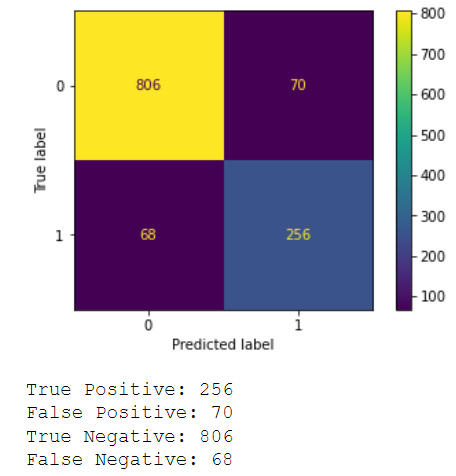


Рисунок 3 – Значения матрицы ошибок для метода k-ближайших соседей

Разработаем предсказательную модель качественного отклика методом логистической регрессии и подсчитаем метрики "Accuracy", "Precision", "Recall", "Balanced accuracy", "F1 score" (Рисунок 4):

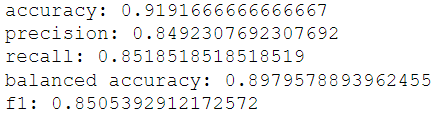


Рисунок 4 – Метрики оценки классификации методом логистической регрессии

Высчитаем значения матрицы ошибок для метода логистической регрессии (Рисунок 5):

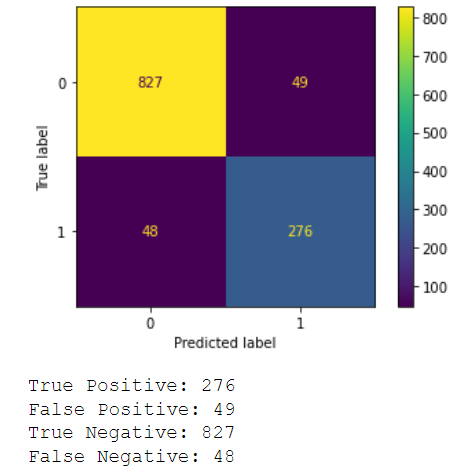


Рисунок 5 – Значения матрицы ошибок для метода логистической регрессии

Разработаем предсказательную модель качественного отклика методом случайного леса и подсчитаем метрики "Accuracy", "Precision", "Recall", "Balanced accuracy", "F1 score" (Рисунок 6):

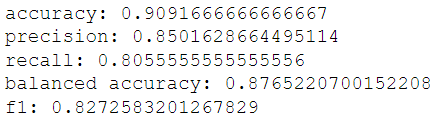


Рисунок 6 – Метрики оценки классификации методом случайного леса

Высчитаем значения матрицы ошибок для метода случайного леса (Рисунок 7):

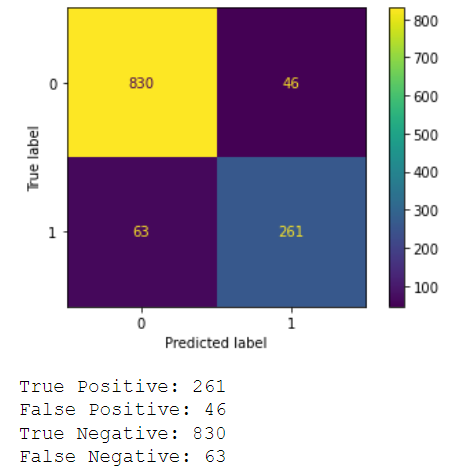


Рисунок 7 – Значения матрицы ошибок для метода случайного леса

Построим графики ROC-кривых (Рисунок 8):

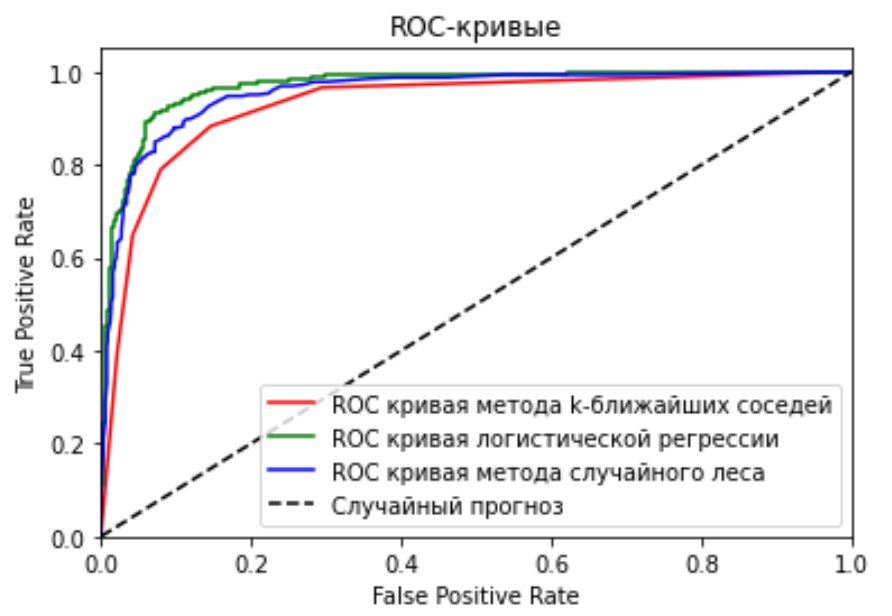


Рисунок 8 - Графики ROC-кривых

Ссылка на GitHub с проектом:

<https://github.com/DmitryPotemkin/VvAD_Labs/blob/main/VvADLab4.ipynb>

1. **Вывод:** в ходе выполнения лабораторной работы изучены алгоритмы и методы классификации данных. Были разработаны предсказательные модели, вычисляющие по набору признаков, уйдёт ли клиент из фитнес-клуба на основе методов: k- ближайших соседей, логистической регрессии, случайного леса. выли высчитаны метрики, позволяющие оценить качество работы классификаторов. Лучше всего высчитывает предсказания классификатор, основанный на методе логистической регрессии, а хуже всех себя показал метод k-ближайших соседей. Тем не менее, все три метода классификации данных показали неплохие результаты. Для каждого метода были, также, были построены матрицы неточностей и график ROC-кривых, который подтвердил полученные оценки качества работы методов классификации.