|  |  |
| --- | --- |
|  | **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 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Отклик клиентов на\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_рекламные предложения\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_\_ИУ5-64Б\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_Ковыршин П.А.\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Гапанюк Ю.Е.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_ Гапанюк Ю.Е.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2023 г.*

**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)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_Отклик клиентов на рекламные предложения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_ИУ5-64Б\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Ковыршин Павел Алексеевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Исследоваткльская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к \_\_\_ нед., 50% к \_\_\_ нед., 75% к \_\_ нед., 100% к \_\_\_ нед.

Техническое задание \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_Исследовать методы машинного обучения для решения задачи классификации\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \_\_32\_\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания « 13 » февраля 2023 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Гапанюк Ю.Е.\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Ковыршин П.А.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Содержание**

[**Введение** 4](#_Toc133806658)

[**Постановка задачи** 6](#_Toc133806659)

[**Выполнение работы** 7](#_Toc133806660)

[**Создание веб-приложения** 29](#_Toc133806661)

[**Заключение** 31](#_Toc133806662)

[**Список использованной литературы** 32](#_Toc133806663)

# **Введение**

Каждый бизнес рано или поздно сталкивается с необходимостью рекламы. Одна из стратегий – специальные предложения клиентам. Однако реклама стоит больших денег, потому для большей прибыли целесообразно искать способы повысить эффективность рекламных кампаний путем анализа клиентской базы и поиска клиентов, которые вероятнее всего откликнутся на предложение.

В данной работе будут использоваться данные о клиентах и их отклике на рекламные предложения, чтобы построить модель машинного обучения, которая сможет заранее предсказать, купил ли данный клиент предложенный в рамках рекламной кампании продукт. Для предсказаний будут использоваться алгоритмы классификации вместе с набором данных, содержащим информацию о клиенте и его предыдущем взаимодействии с компанией.

Целью данной работы является разработка модели, которая сможет определять клиентов, охотно откликающихся на рекламные предложения, чтобы помочь бизнесу сэкономить время и ресурсы, затрачиваемые на рекламу.

Для достижения поставленной цели были определены следующие этапы:

1. Поиск и выбор набора данных для построения моделей машинного обучения для решения задачи классификации.
2. Проведение разведочного анализа данных.
3. Выбор признаков, подходящих для построения моделей.
4. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
5. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
6. Выбор метрик для последующей оценки качества моделей.
7. Выбор наиболее подходящих моделей для решения задачи классификации.
8. Формирование обучающей и тестовой выборок на основе исходного набора данных.
9. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров и оценка качества моделей на основе тестовой выборки.
10. Подбор гиперпараметров для выбранных моделей. Построение оптимальных моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.
12. Создание веб-приложение для демонстрации хотя бы одной модели машинного обучения.

# **Постановка задачи**

Данная работа по машинному обучению направлена на решение задачи классификации, а именно предсказания, откликнется ли клиент на рекламное предложение.

Имеются данные о клиентах, которые включают информацию такую, как год рождения, уровень образования, семейное положение, доход, количество детей, дата регистрации в компании, время, прошедшее с последней покупки, информация о жалобах и покупках товаров разными способами, а также суммы, потраченные на различные виды товаров, и информация о об отклике на конкретную рекламную кампанию.

Целью – создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания того, воспользуется ли клиент рекламным предложением. Для этого будут использоваться различные алгоритмы классификации, такие как метод K ближайших соседей, логистическая регрессия, метод опорных векторов, случайный лес и градиентный бустинг. Модель должна обучаться на тренировочных данных и проверяться на тестовых данных для оценки ее точности и эффективности.

Результатом работы должна быть модель, которая сможет определять клиентов, которые откликаются на рекламные предложения с приемлемой точностью.

# **Выполнение работы**

Для решения задачи классификации был выбран набор данных, содержащий информацию о клиентах компаний.

В наборе данных присутствуют следующие столбцы:

* *ID* – уникальный идентификатор
* *Year\_Birth* – год рождения клиента
* *Education* – уровень образования клиента
* *Marital\_Status* – семейное положение клиента
* *Income* – доход клиента
* *Kidhome* – количество малолетних детей в доме клиента
* *Teenhome* – количество детей-подростков в доме клиента
* *Dt\_Customer* – дата первого регистрации клиента в компании
* *Recency* – количество дней, прошедших с последней покупки клиента
* *Complain* – жаловался ли клиент в течение последних 2-х лет
* *MntWines* – сумма, потраченная клиентом на вино за последние 2 года
* *MntFruits* – сумма, потраченная клиентом на фрукты за последние 2 года
* *MntMeatProducts* – сумма, потраченная клиентом на мясо за последние 2 года
* *MntFishProducts* – сумма, потраченная клиентом на рыбу за последние 2 года
* *MntSweetProducts* – сумма, потраченная клиентом на сладости за последние 2 года
* *MntGoldProducts* – сумма, потраченная клиентом на золото за последние 2 года
* *NumDealsPurchases* – количество покупок клиента, сделанных со скидкой
* *AcceptedCmp1* – принял ли клиент предложение во время первой кампании
* *AcceptedCmp2* – принял ли клиент предложение во время второй кампании
* *AcceptedCmp3* – принял ли клиент предложение во время третьей кампании
* *AcceptedCmp4* – принял ли клиент предложение во время четвертой кампании
* *AcceptedCmp5* – принял ли клиент предложение во время пятой кампании
* *Response* – принял ли клиент предложение во время последней кампании
* *NumWebPurchases* – количество покупок клиента, сделанных через вебсайт
* *NumCatalogPurchases* – количество покупок клиента, сделанных через каталог
* *NumStorePurchases* – количество покупок клиента, сделанных напрямую в магазине
* *NumWebVisitsMonth* – количество посещений клиентом вебсайта за последний месяц

Данный набор данных был использован для решения задачи классификации – предсказания, откликнется ли клиент на рекламное предложение.

Загружаем набор данных, очищая от возможных дубликатов, и получаем общую информацию о нем. В наборе данных содержится 2240 строк и 29 столбцов, из которых 3 строковых, 1 типа float64 и 25 целочисленных.

Обнаруживаем 2 столбца, не указанных в описании набора данных: Z\_CostContact и Z\_Revenue. После проверки оказывается, что каждый из них содержит единственное значение во всех строках, поэтому заранее удаляем их.

В наборе данных также содержатся уникальные идентификаторы ID, которые очевидно не несут в себе полезной информации для решения поставленной задачи.

Так как целью работы является определение того, откликнулся ли клиент на рекламное предложение в принципе, объединяем столбцы AcceptedCmp[1-5] и Response в новый столбец Bought, записывая в него 1, если в каком-либо из исходных столбцов содержалось значение 1, иначе записывая 0.

Столбец Dt\_Customer, содержащий в себе дату регистрации в компании, заменим на числовой, записав в него условный номер дня с 01.01.1970.

Проверим набор данных на наличие пустых значений.

Обнаружили 24 строки с пустым значением Income – удаляем.

Строим график pairplot для визуализации попарного распределения данных для множества колонок, отображая цветом разные значения целевого признака.

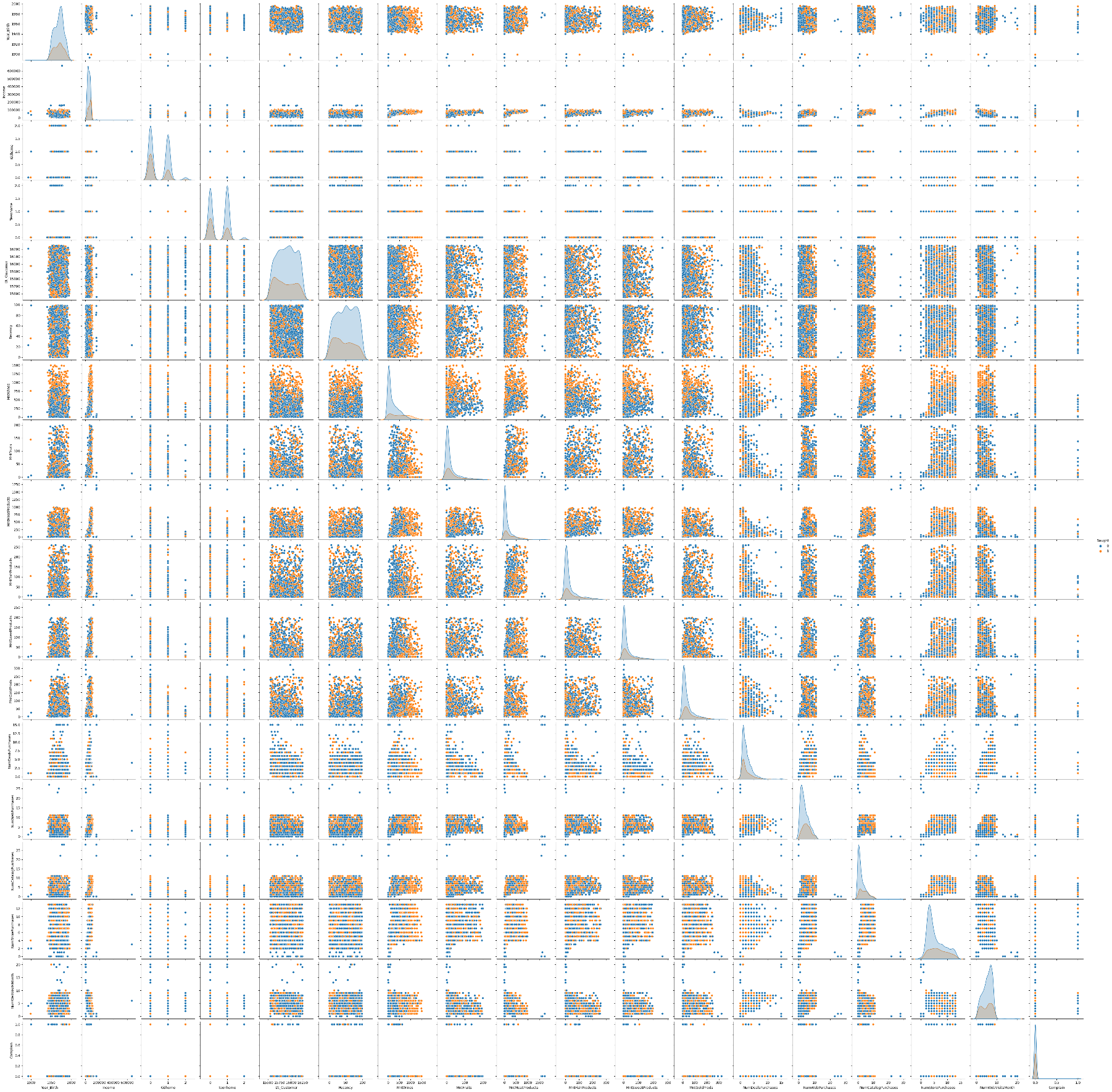


Рисунок 1 - Визуализация попарного распределения данных для множества колонок

Проверяем, сбалансированы ли классы целевого признака в наборе данных. Получаем следующую гистограмму:

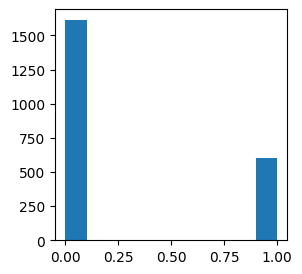


Рисунок 2 - Гистограмма классов

Видим, что дисбаланс классов присутствует (0: 72.7%, 1: 27.3%), но является приемлемым.

Строим скрипичные диаграммы для всех числовых признаков, чтобы оценить примерное их распределение.

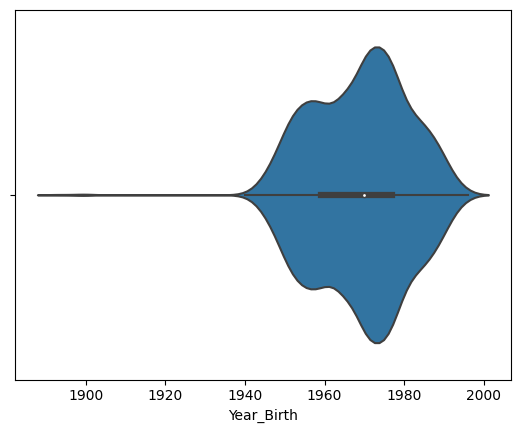


Рисунок 3 – Скрипичная диаграмма распределения годов рождения клиентов

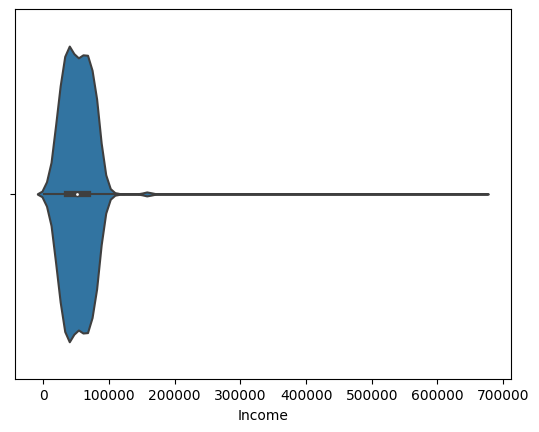


Рисунок 4 – Скрипичная диаграмма распределения доходов клиентов

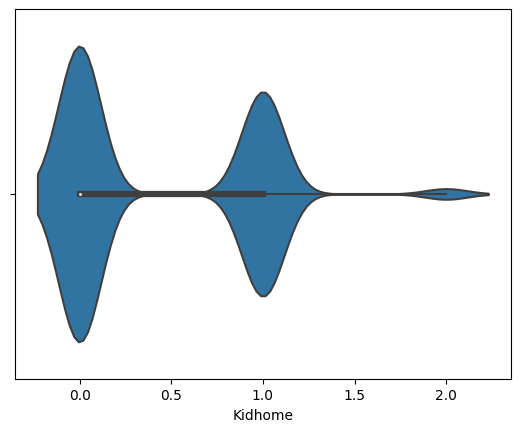


Рисунок 5 – Скрипичная диаграмма распределения количества малолетних детей в домах клиентов

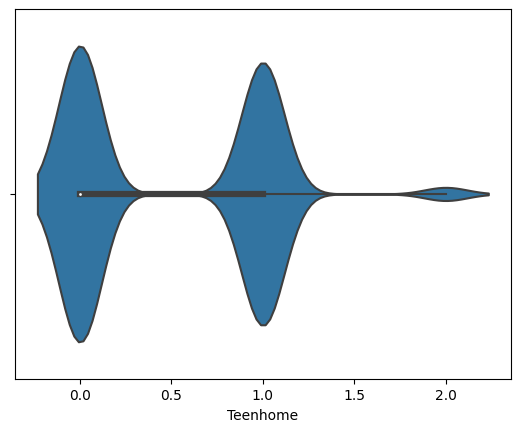


Рисунок 6 – Скрипичная диаграмма распределения количества детей подросткового возраста в домах клиентов

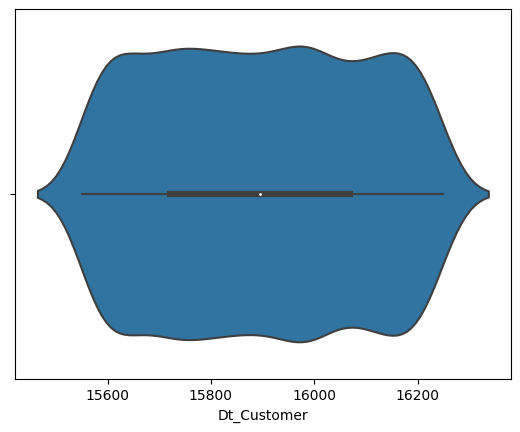


Рисунок 7 – Скрипичная диаграмма распределения дат регистрации клиентов в компании

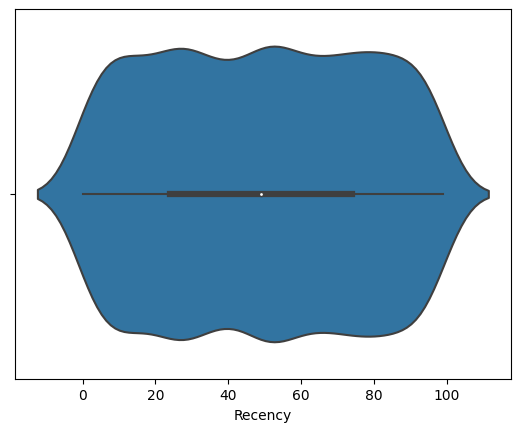


Рисунок 8 – Скрипичная диаграмма распределения количества дней, прошедших с последней покупки клиента

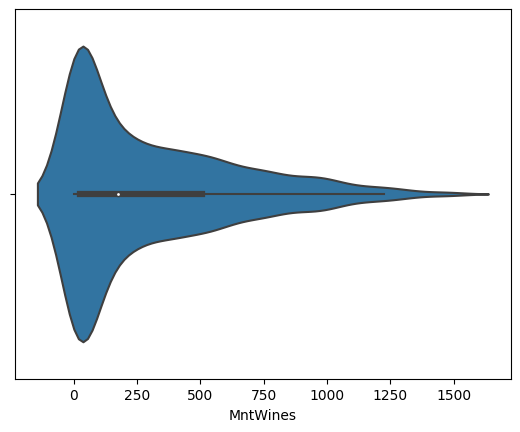


Рисунок 9 – Скрипичная диаграмма распределения сумм, потраченных клиентами на вино

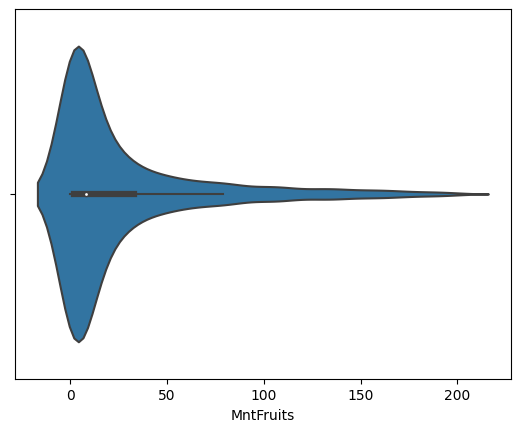


Рисунок 10 – Скрипичная диаграмма распределения сумм, потраченных клиентами на фрукты

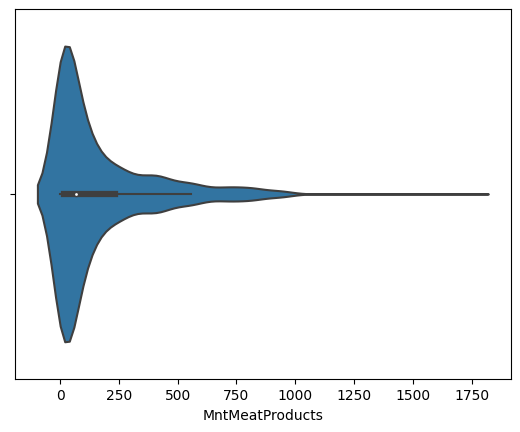


Рисунок 11 – Скрипичная диаграмма распределения сумм, потраченных клиентами на мясо

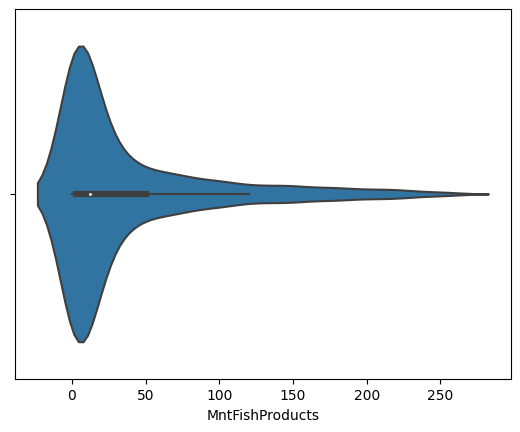


Рисунок 12 – Скрипичная диаграмма распределения сумм, потраченных клиентами на рыбу

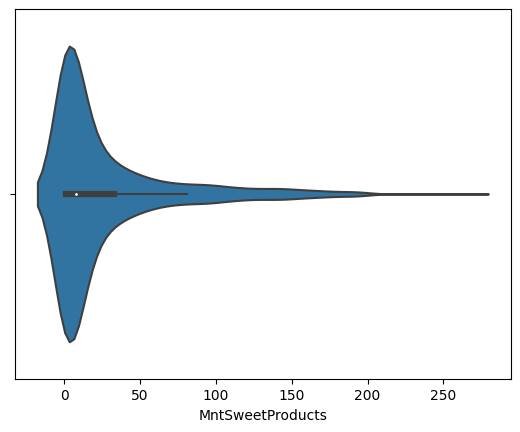


Рисунок 13 – Скрипичная диаграмма распределения сумм, потраченных клиентами на сладости

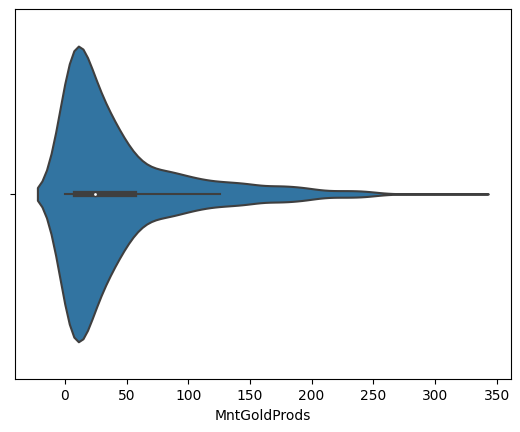


Рисунок 14 – Скрипичная диаграмма распределения сумм, потраченных клиентами на золото

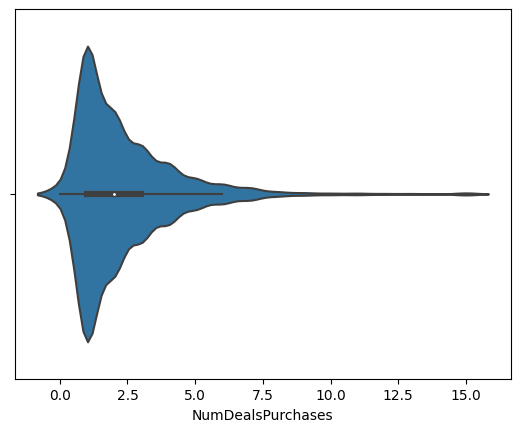


Рисунок 15 – Скрипичная диаграмма распределения количества покупок клиентов, сделанных со скидкой

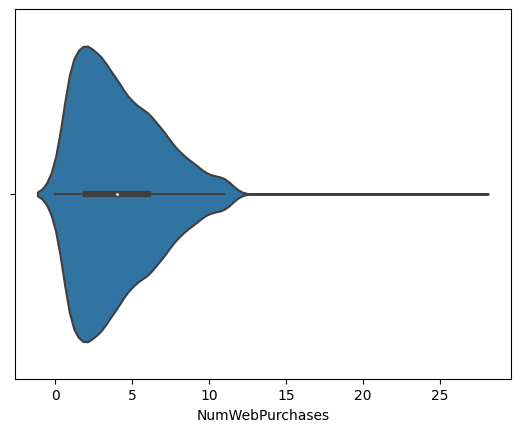


Рисунок 16 – Скрипичная диаграмма распределения количества покупок клиентов, сделанных через вебсайт

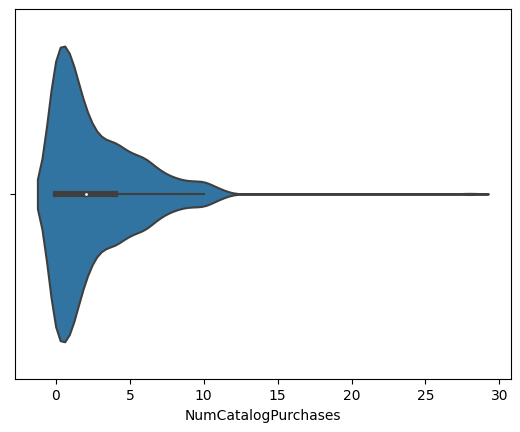


Рисунок 17 – Скрипичная диаграмма распределения количества покупок клиентов, сделанных через каталог

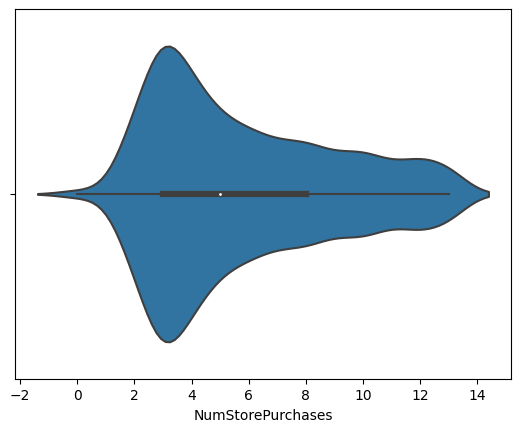


Рисунок 18 – Скрипичная диаграмма распределения количества покупок клиентов, сделанных напрямую в магазине

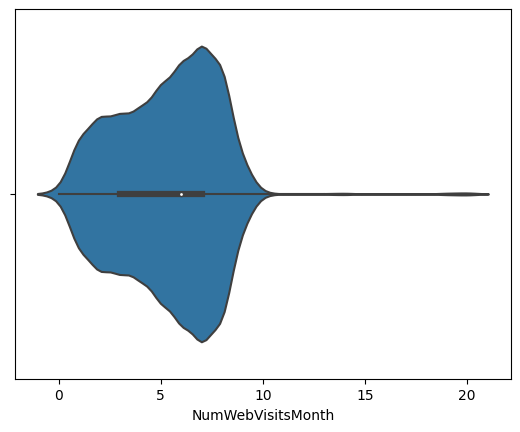


Рисунок 19 – Скрипичная диаграмма распределения количества посещений клиентом вебсайта

Приведем данные в удобную для моделей машинного обучения форму.

Масштабируем числовые признаки методом MinMax, приводя их к промежутку [0;1] с сохранением распределения. По гистограммам убедимся, что масштабирование не повлияло на распределение значений.

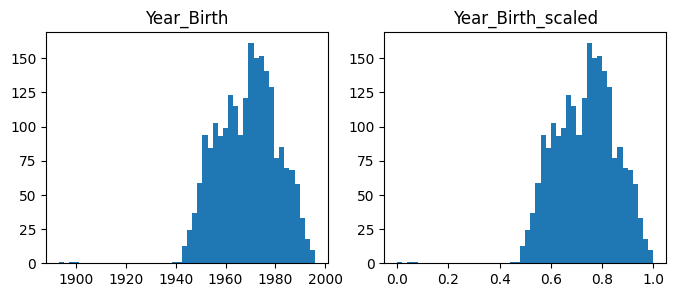


Рисунок 20 – Распределение колонки Year\_Birth до и после масштабирования

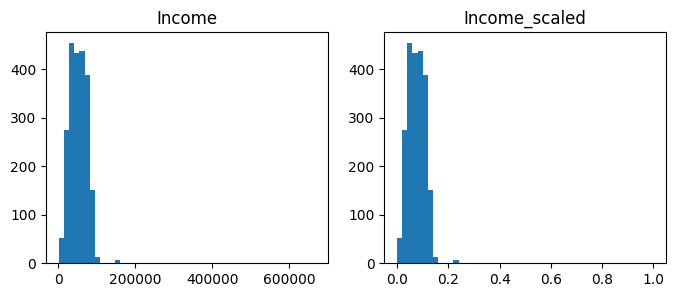


Рисунок 21 – Распределение колонки Income до и после масштабирования

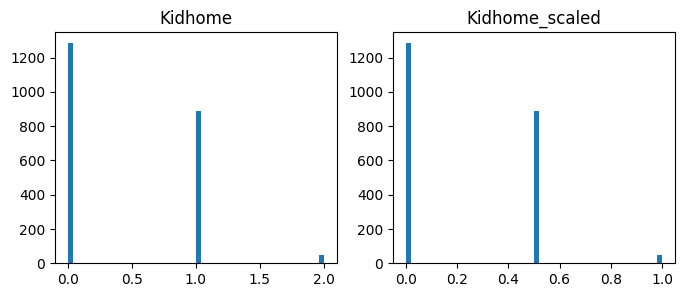


Рисунок 22 – Распределение колонки Kidhome до и после масштабирования

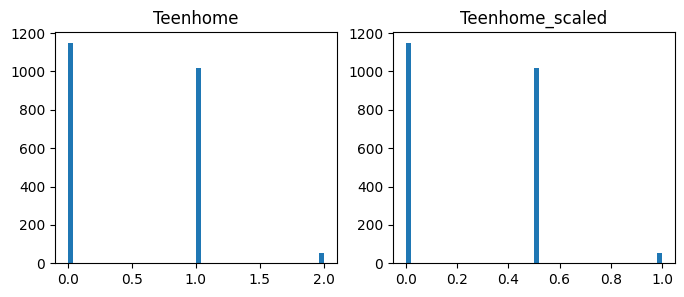


Рисунок 23 – Распределение колонки Teenhome до и после масштабирования

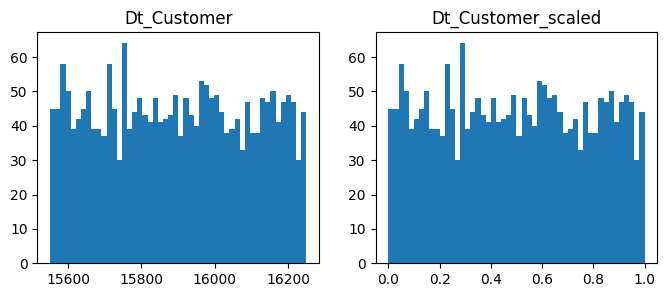


Рисунок 24 – Распределение колонки Dt\_Customer до и после масштабирования

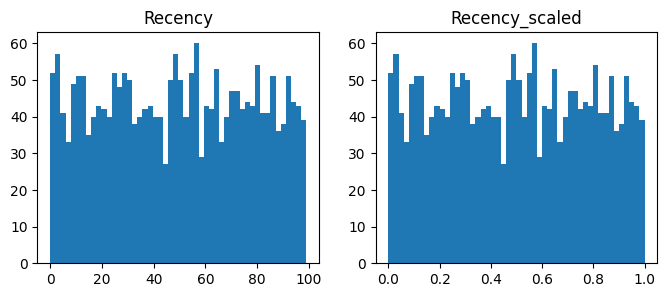


Рисунок 25 – Распределение колонки Recency до и после масштабирования

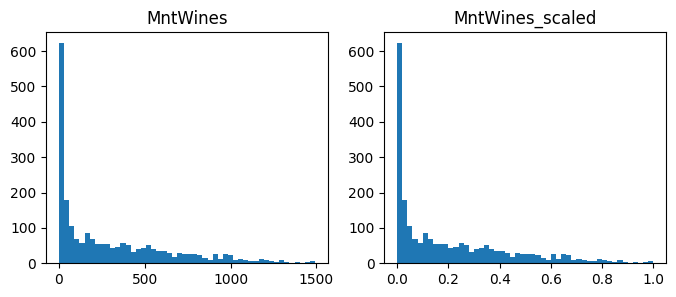


Рисунок 26 – Распределение колонки MntWines до и после масштабирования

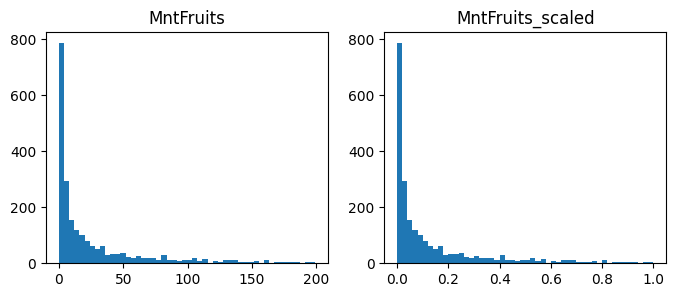


Рисунок 27 – Распределение колонки MntFruits до и после масштабирования

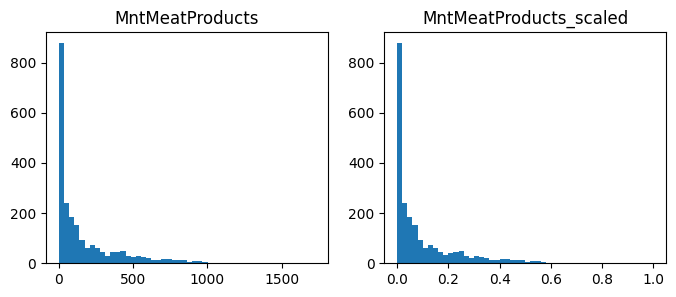


Рисунок 28 – Распределение колонки MntMeatProducts до и после масштабирования

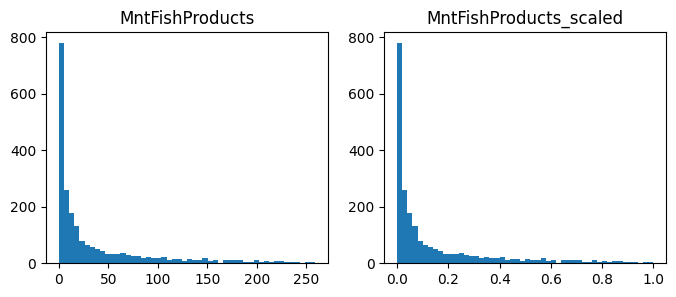


Рисунок 29 – Распределение колонки MntFishProducts до и после масштабирования

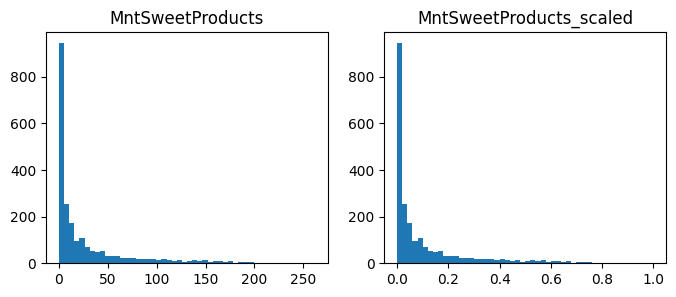


Рисунок 30 – Распределение колонки MntSweetProducts до и после масштабирования

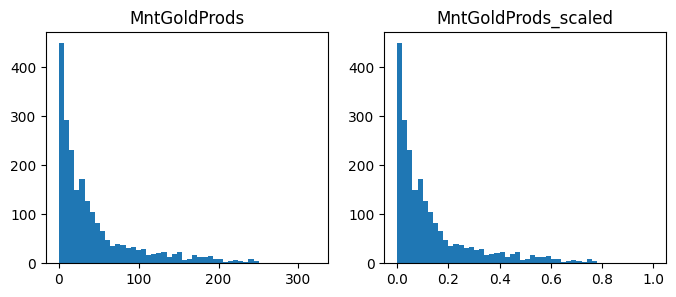


Рисунок 31 – Распределение колонки MntGoldProducts до и после масштабирования

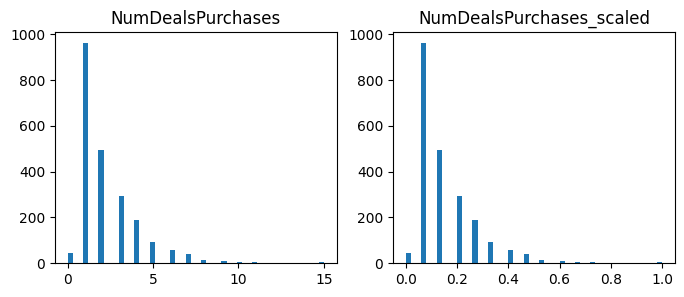


Рисунок 32 – Распределение колонки NumDealsPurchases до и после масштабирования

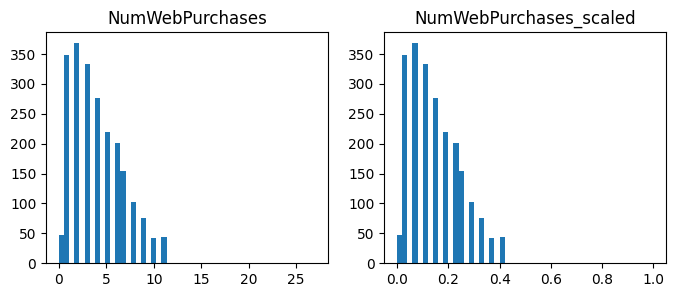


Рисунок 33 – Распределение колонки NumWebPurchases до и после масштабирования

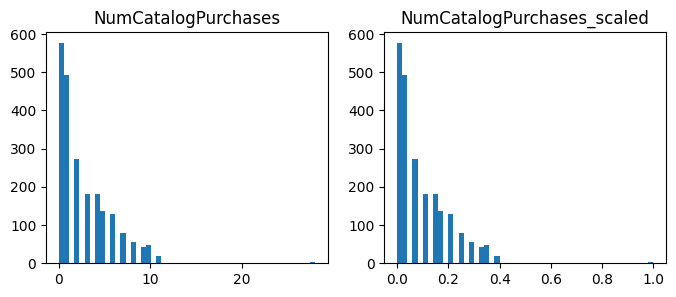


Рисунок 34 – Распределение колонки NumCatalogPurchases до и после масштабирования

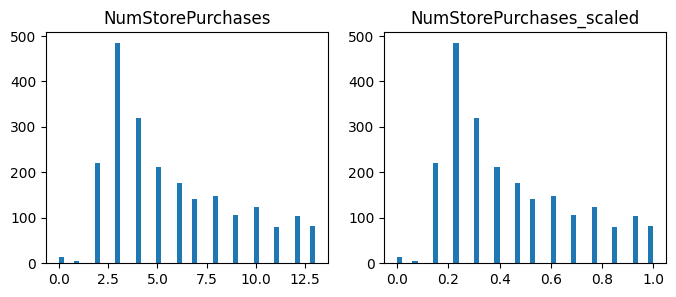


Рисунок 35 – Распределение колонки NumStorePurchases до и после масштабирования

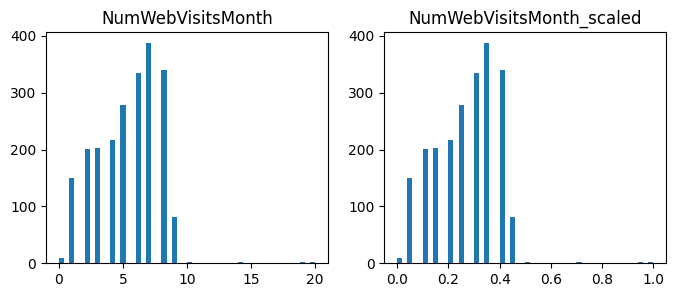


Рисунок 36 – Распределение колонки NumWebVisitsMonth до и после масштабирования

Масштабирование не повлияло на распределение признаков.

Закодируем категориальные признаки, заменяя строковые значения на числа.

Проводим корреляционный анализ данных. Строим тепловую крату корреляций.

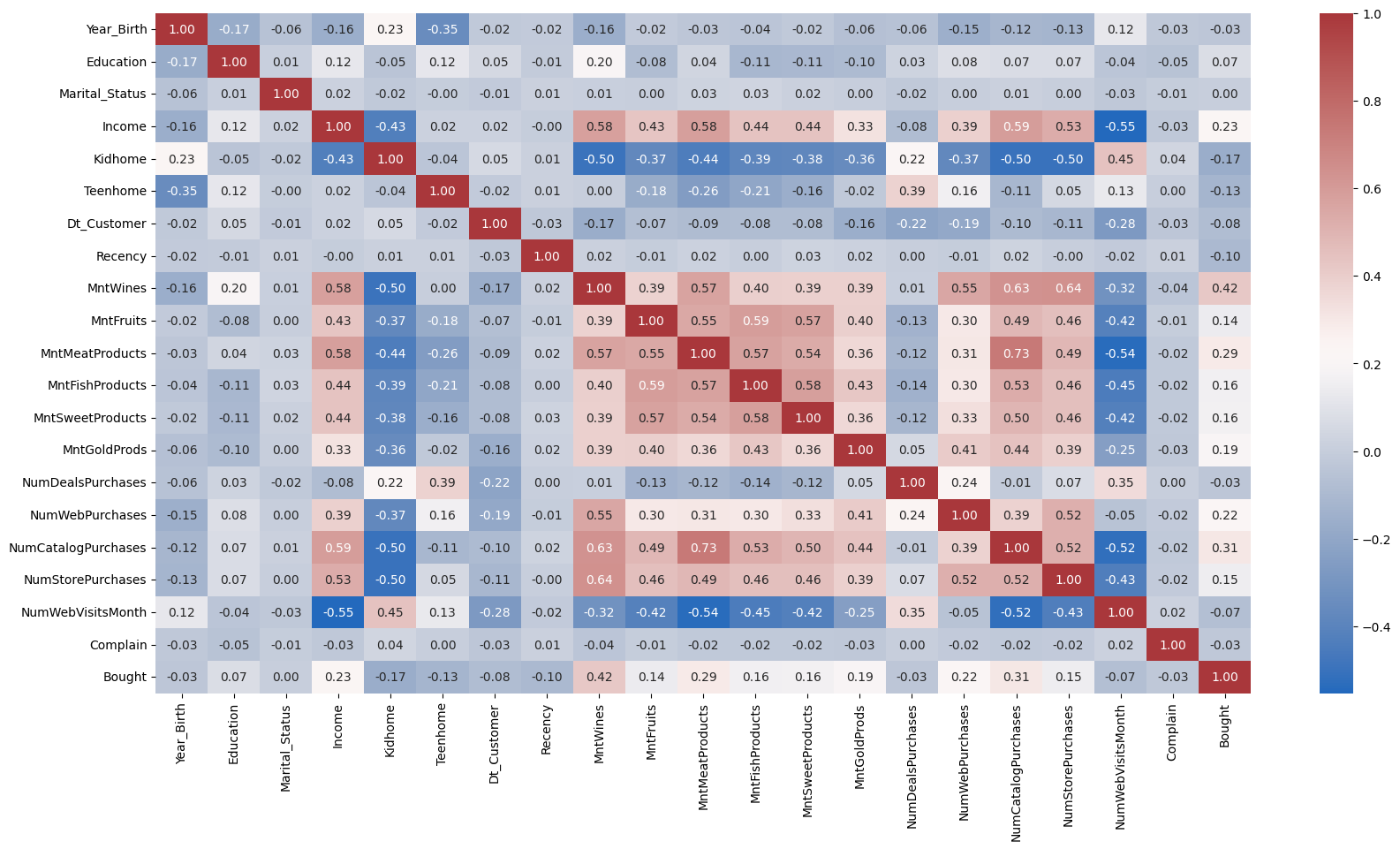


Рисунок 37 – Тепловая карта корреляций

Выводы:

* Целевой признак наиболее сильно коррелирует с MntWines (0.42), NumCatalogPurchases (0.31), MntMeatProducts (0.29). Эти признаки следует оставить в модели.
* Нецелевых признаков, имеющих между собой особо сильную корреляцию, в наборе данных не наблюдается.
* В наборе данных присутствуют признаки, имеющие нулевую корреляцию с целевым. Их можно удалить.
* Все значения коэффициентов корреляции между признаками довольно низкие по модулю, и такие данные, вероятно, нельзя будет использовать для построения модели машинного обучения надлежащего качества.

Удалим признаки, имеющие с целевым низкую по модулю корреляцию (<0.15).

Выберем метрики для оценки качества модели:

* – сбалансированная точность, среднее арифметическое точности для каждого класса по отдельности. [2]
* - среднее гармоническое precision и recall. Другими словами, это средневзвешенное значение точности и отзыва. [3]
* - oснована на вычислении следующих характеристик: - True Positive Rate, откладывается по оси ординат. Совпадает с recall. - False Positive Rate, откладывается по оси абсцисс. Показывает какую долю из объектов отрицательного класса алгоритм предсказал неверно. Идеальная ROC-кривая проходит через точки (0,0)-(0,1)-(1,1), то есть через верхний левый угол графика. Чем сильнее отклоняется кривая от верхнего левого угла графика, тем хуже качество классификации. [3]

Выберем модели для решения задачи классификации:

* Метод k ближайших соседей;
* Логистическую регрессию;
* Метод опорных векторов;
* Случайный лес;
* Градиентный бустинг.

Формируем обучающую и тестовую выборку.

Обучаем базовые модели, выводим для каждой ROC-кривую и матрицу ошибок.

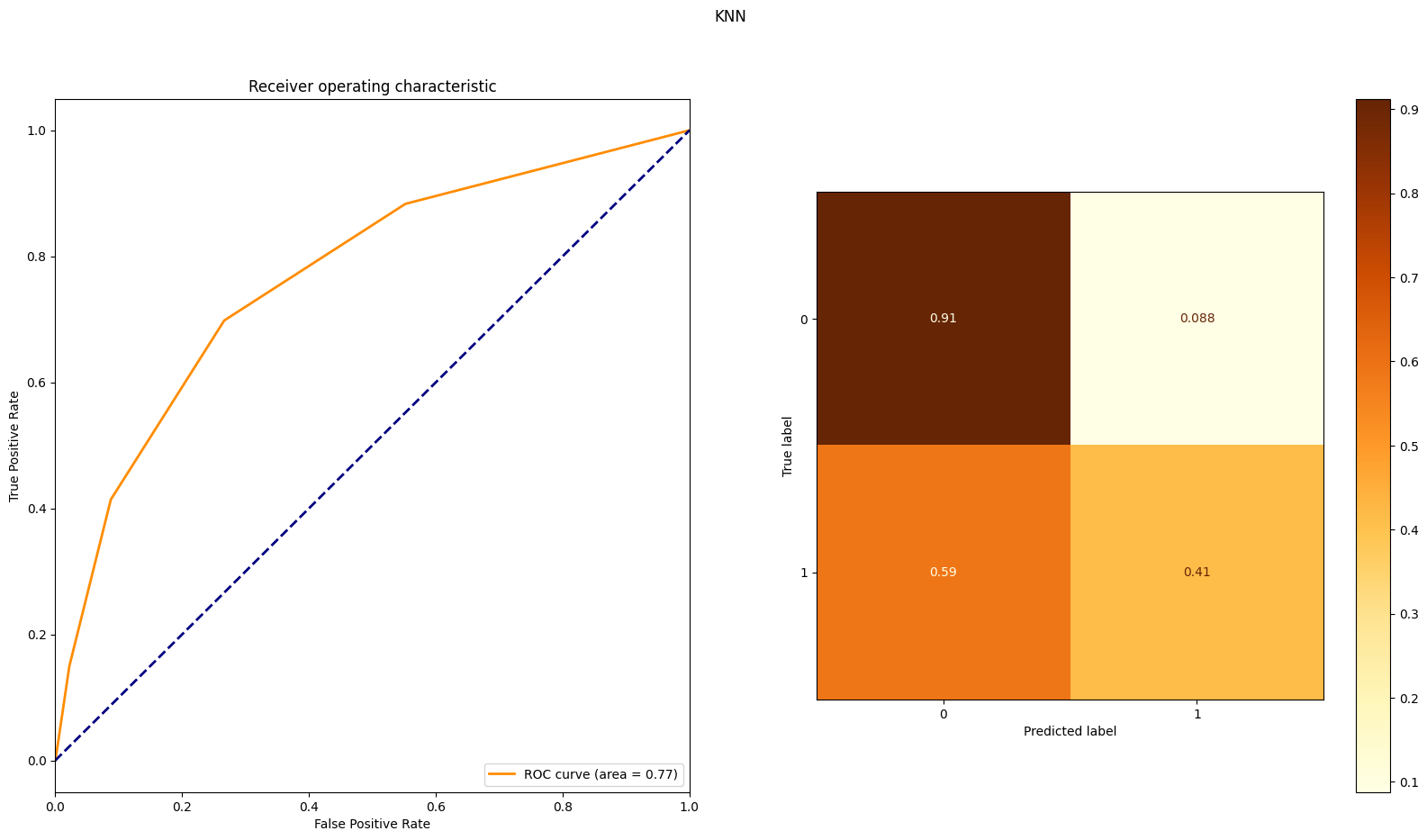


Рисунок 38 – Диаграммы базовой модели метода ближайших соседей

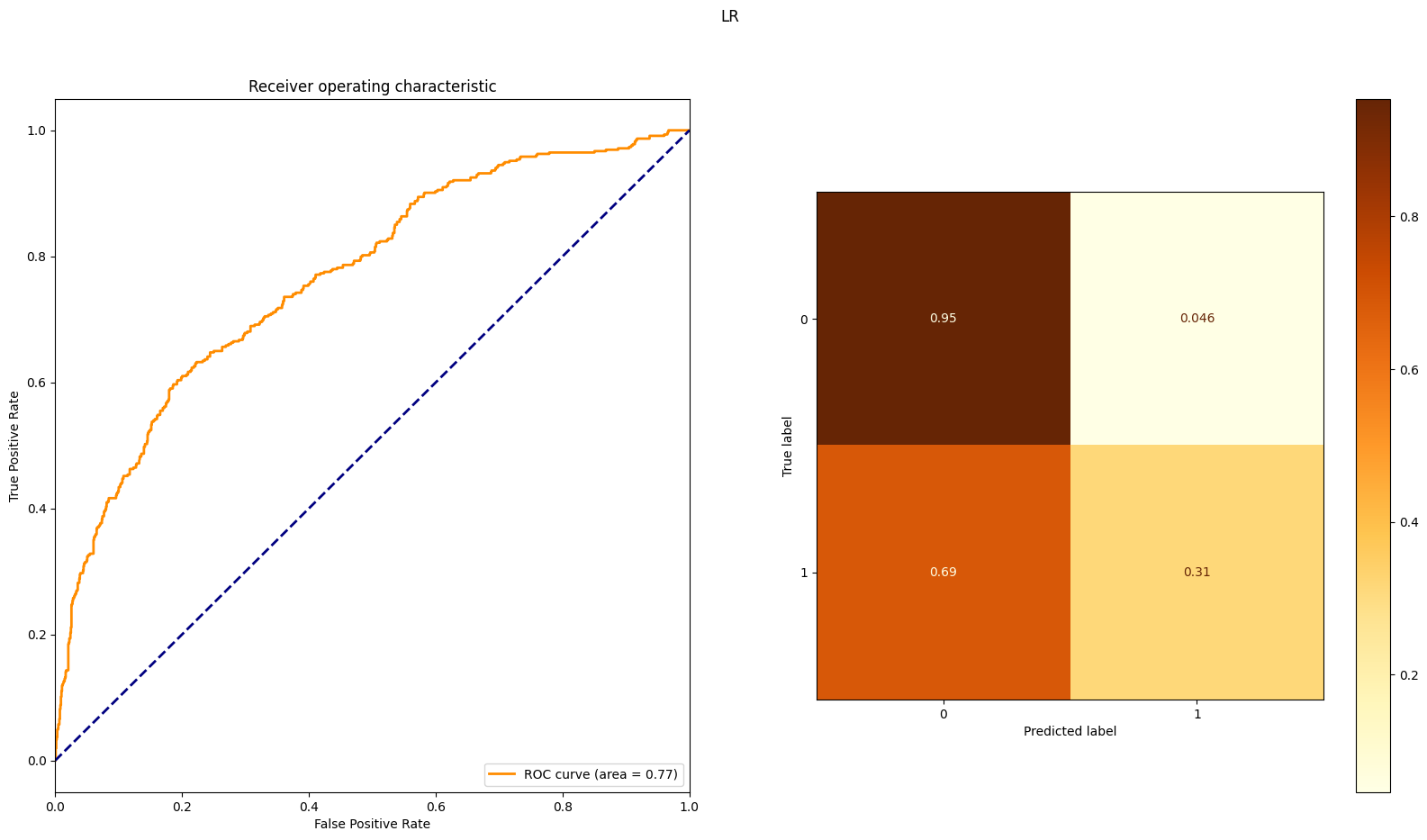


Рисунок 39 – Диаграммы базовой модели логистической регрессии

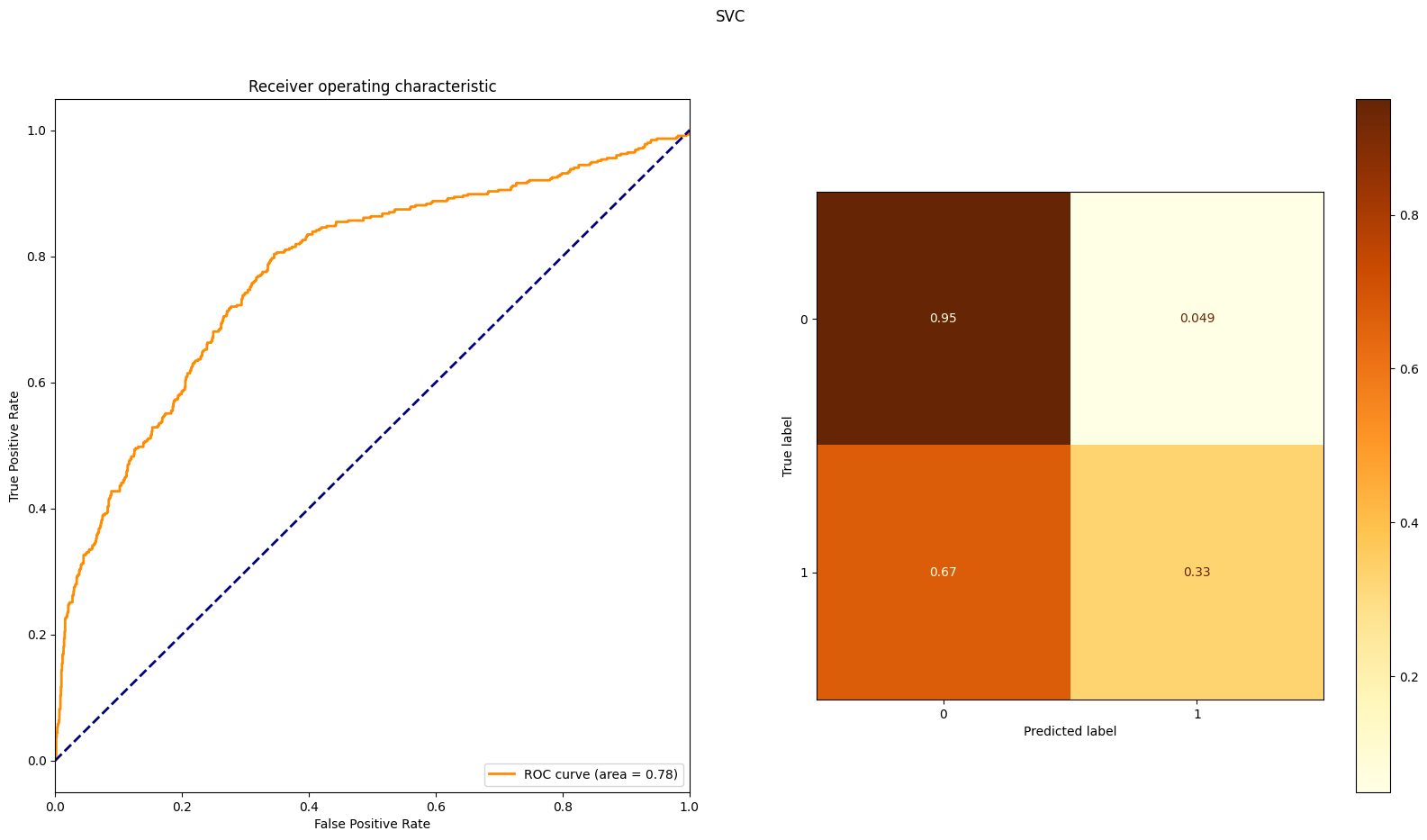


Рисунок 40 – Диаграммы базовой модели метода опорных векторов

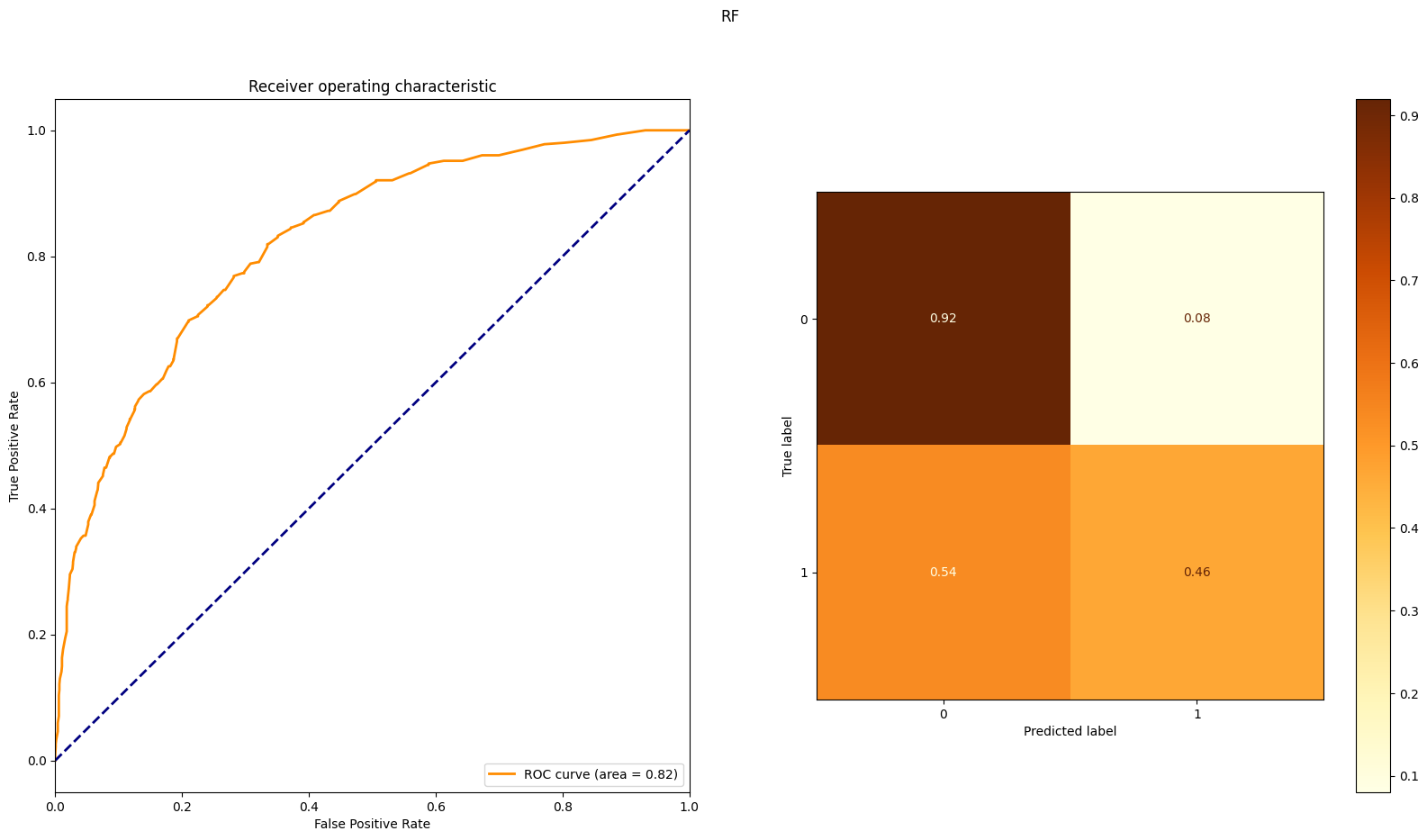


Рисунок 41 – Диаграммы базовой модели случайного леса

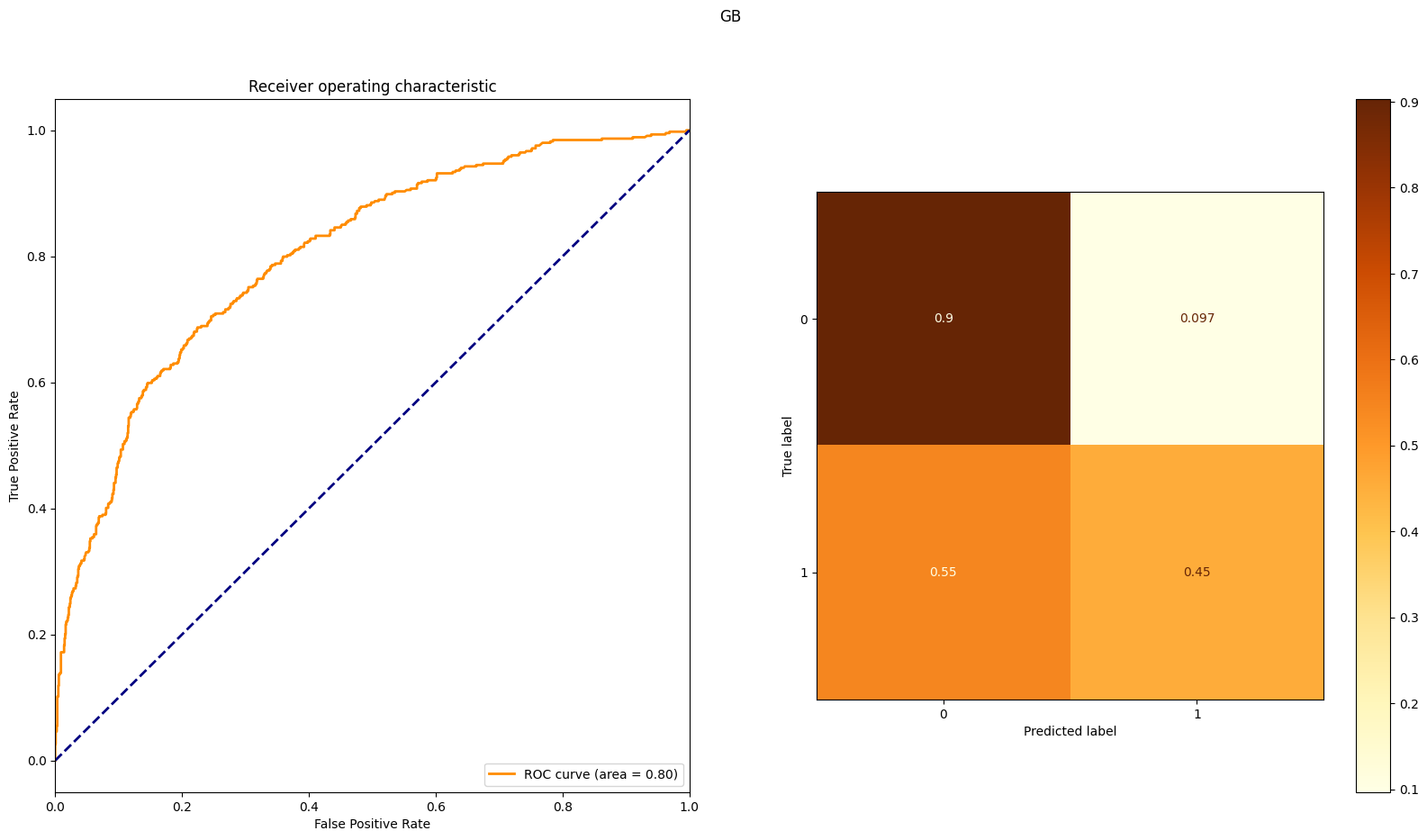


Рисунок 42 – Диаграммы базовой модели градиентного бустинга

Используем RandomSearchCV для поиска оптимальных гиперпараметров для каждой модели.

KNN:

n\_neighbors=12

weights=”distance”

p=2

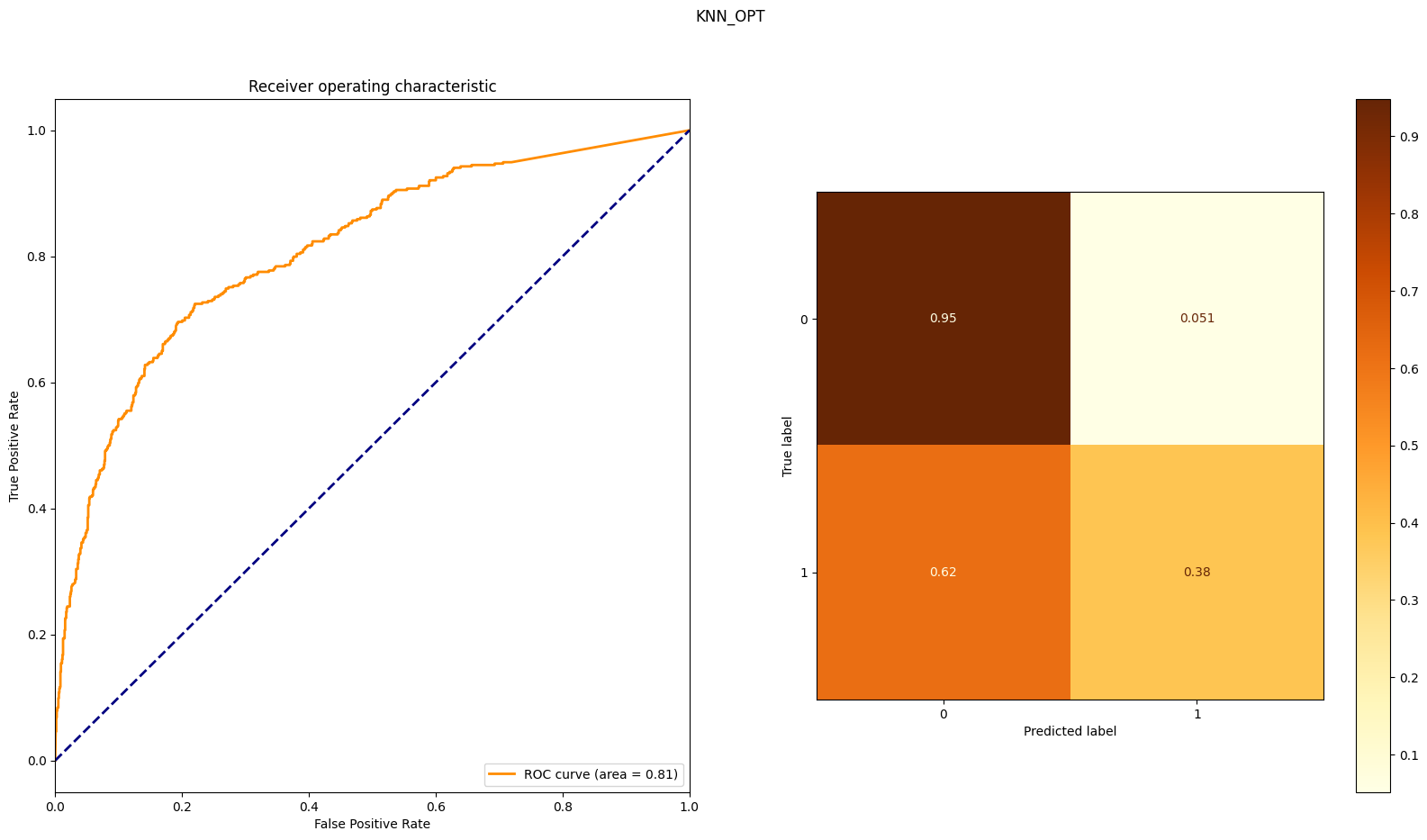


Рисунок 43 – Диаграммы оптимизированной модели метода ближайших соседей

LR:

C=57.543993733715695

solver=”newton-cholesky”

class\_wieght={ 0: 281, 1: 719 }

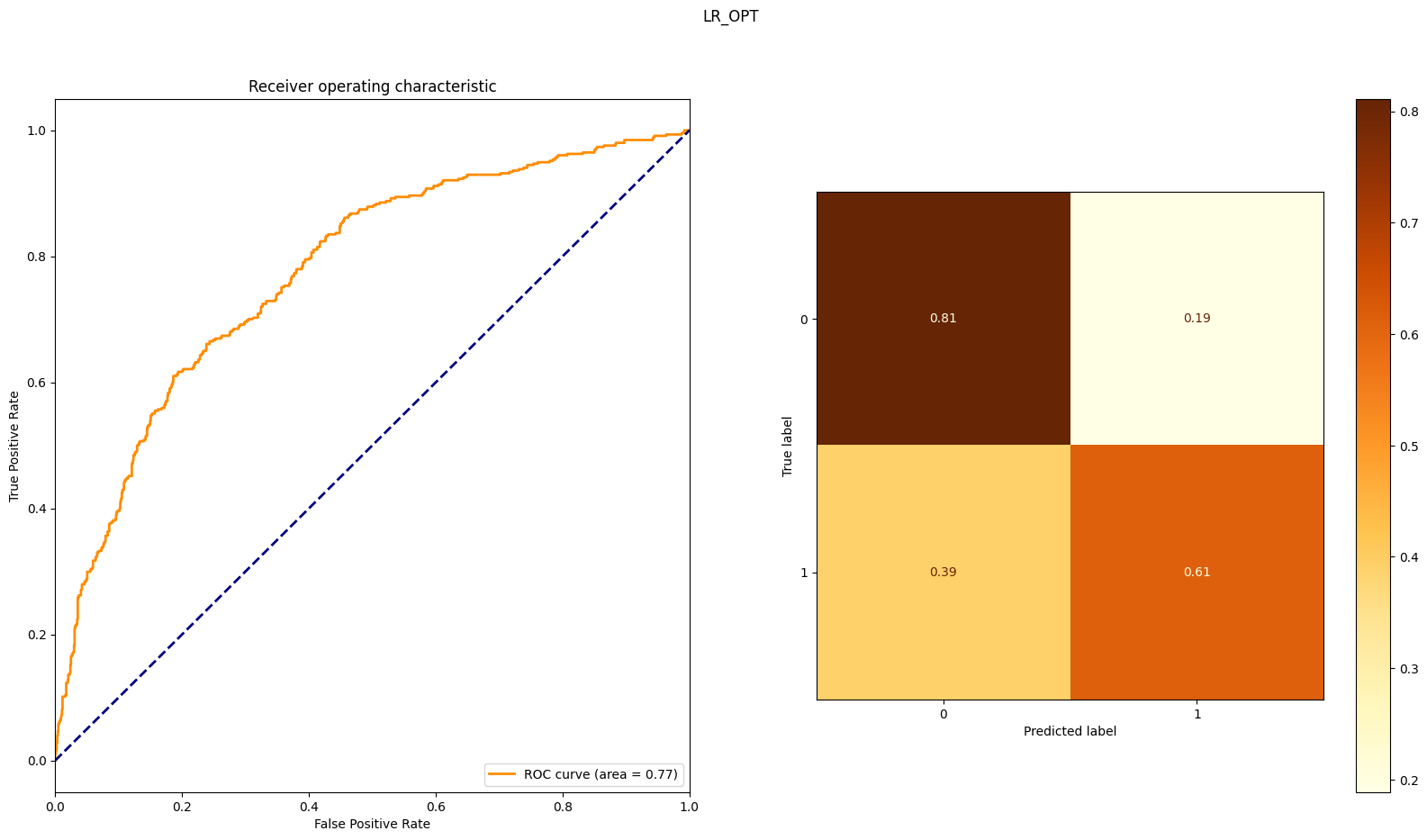


Рисунок 44 – Диаграммы оптимизированной модели логистической регрессии

SVC:

C=0.5011872336272722

kernel=”linear”

class\_weight={ 0: 442, 1: 558 }

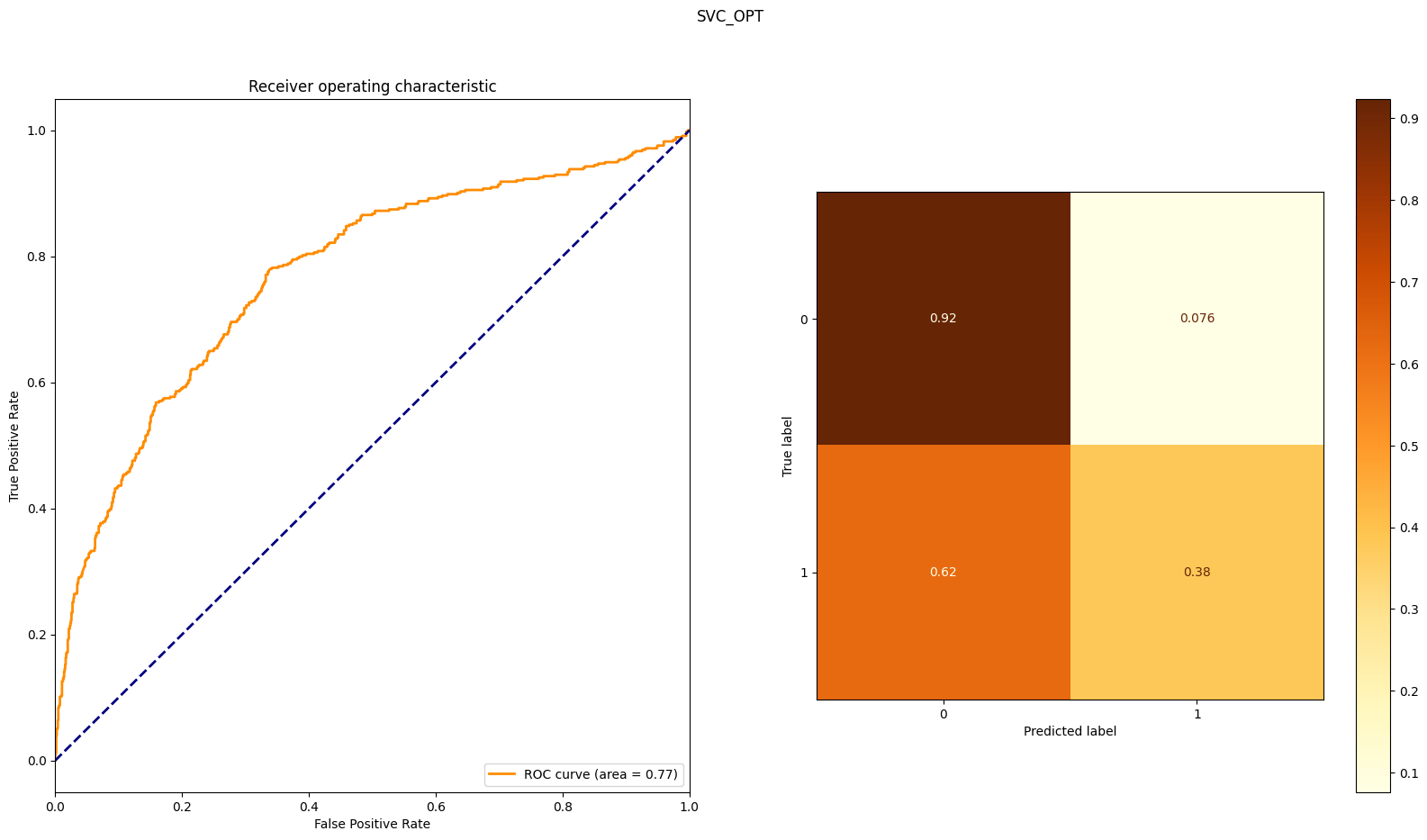


Рисунок 45 – Диаграммы оптимизированной модели метода опорных векторов

RF:

n\_estimators=79

max\_features=”log2”

criterion=”gini”

class\_weight={ 0: 451, 1: 549 }

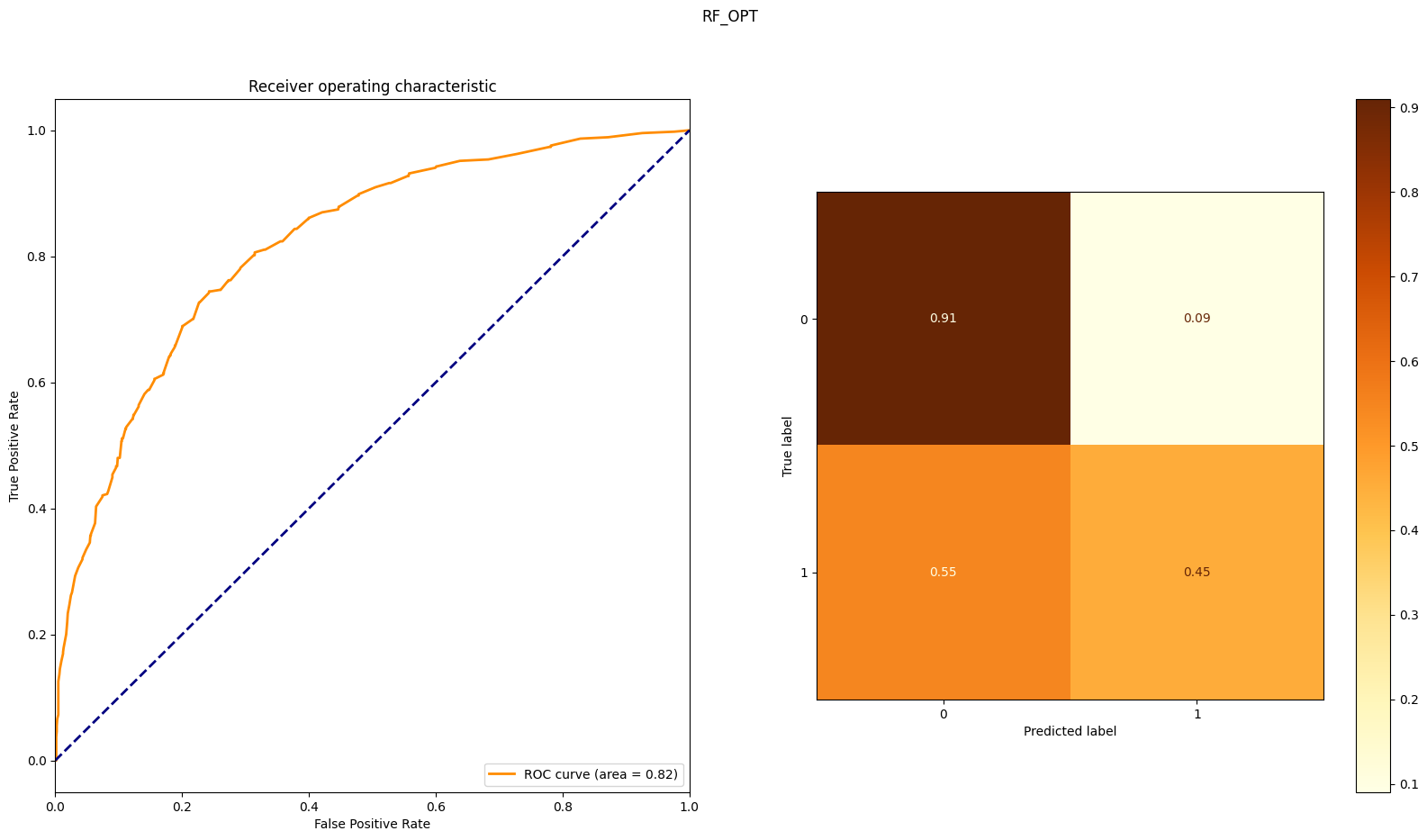


Рисунок 46 – Диаграммы оптимизированной модели случайного леса

GB:

loss=”log\_loss”

learning\_rate=0.07943282347242814

n\_estimators=501

criterion=”squared\_error”

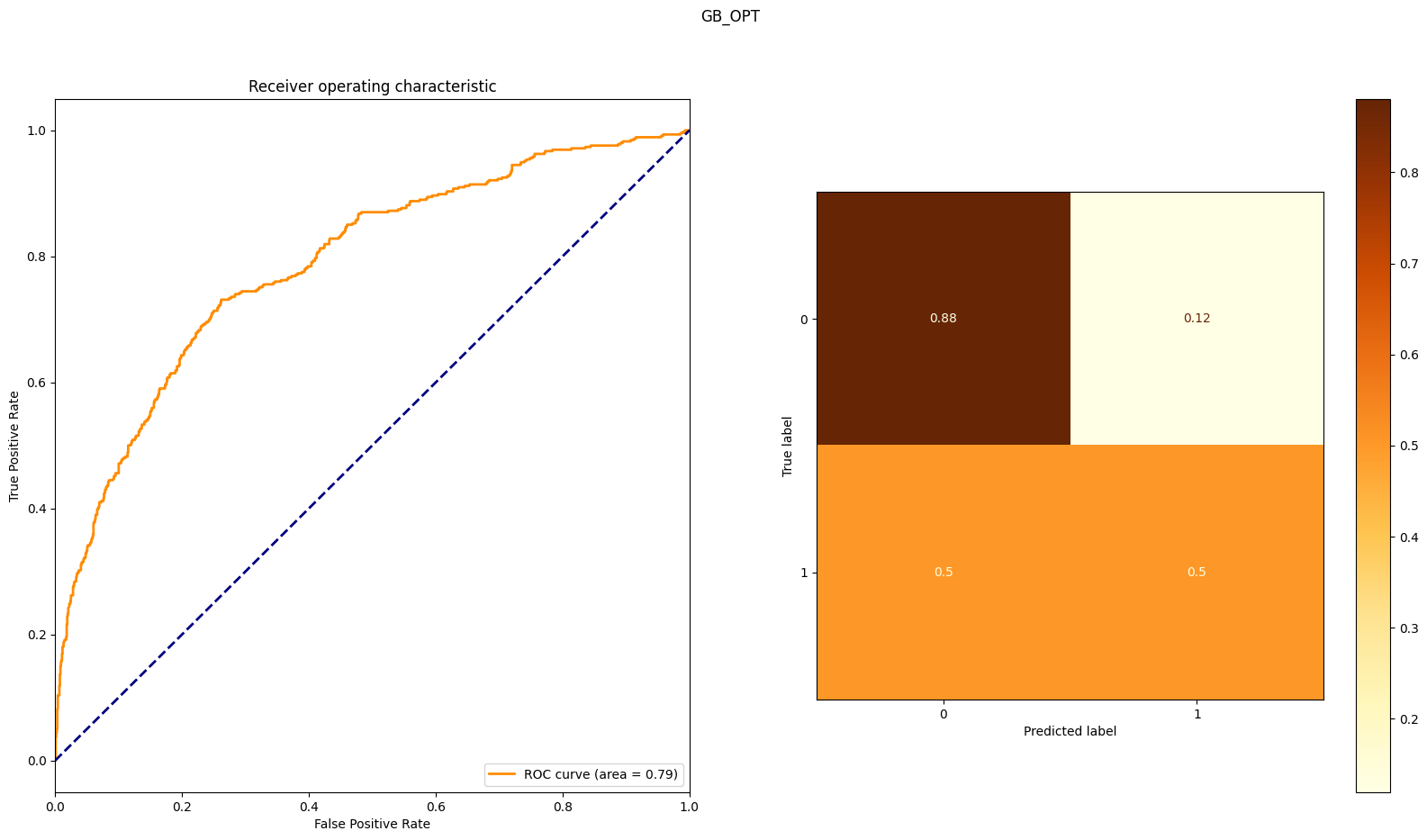


Рисунок 47 – Диаграммы оптимизированной модели градиентного бустинга

Выведем значения для всех метрик, чтобы сравнить все базовые и оптимизированные модели.

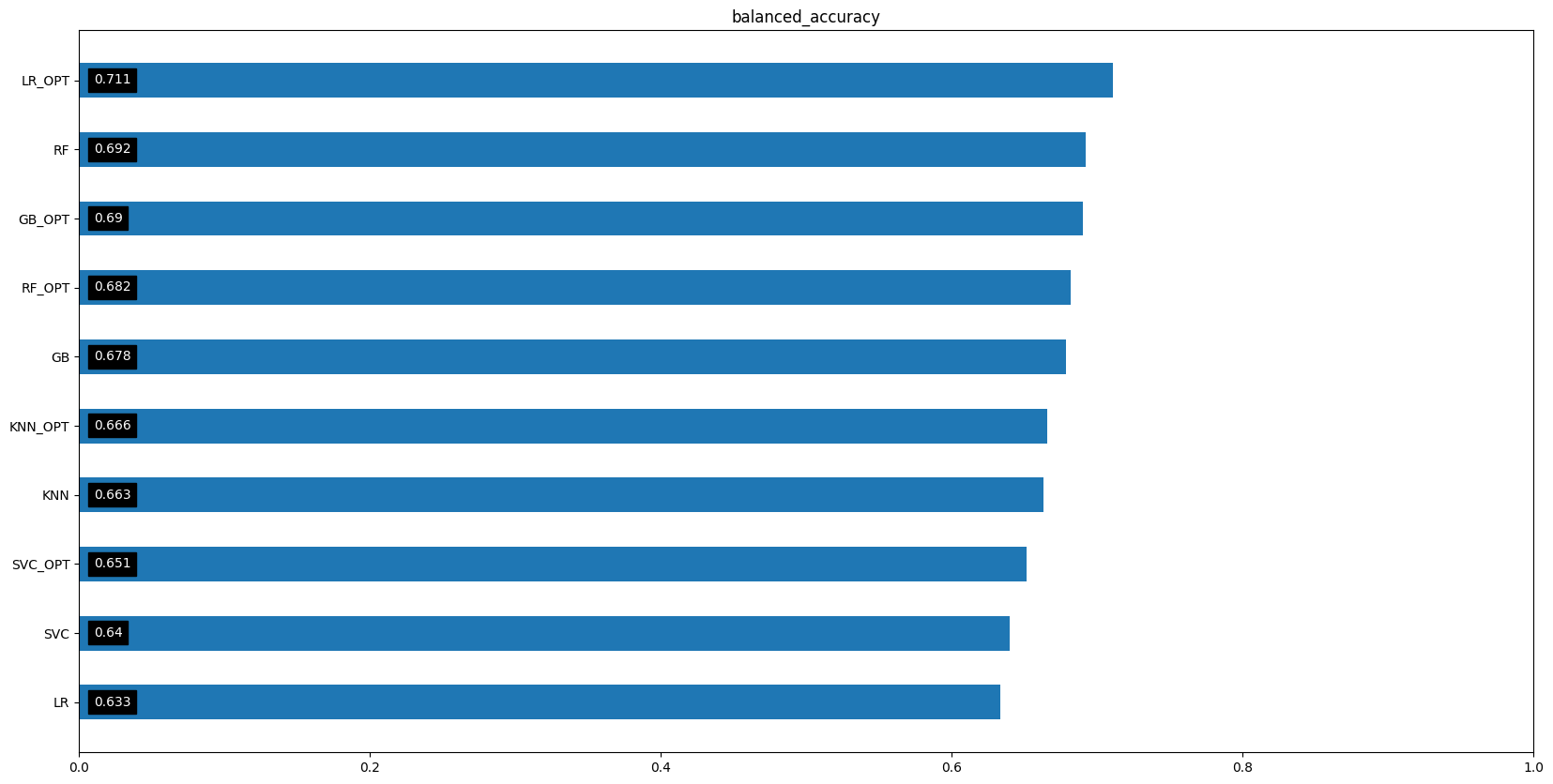


Рисунок 48 – Диаграмма значений метрики balanced\_accuracy

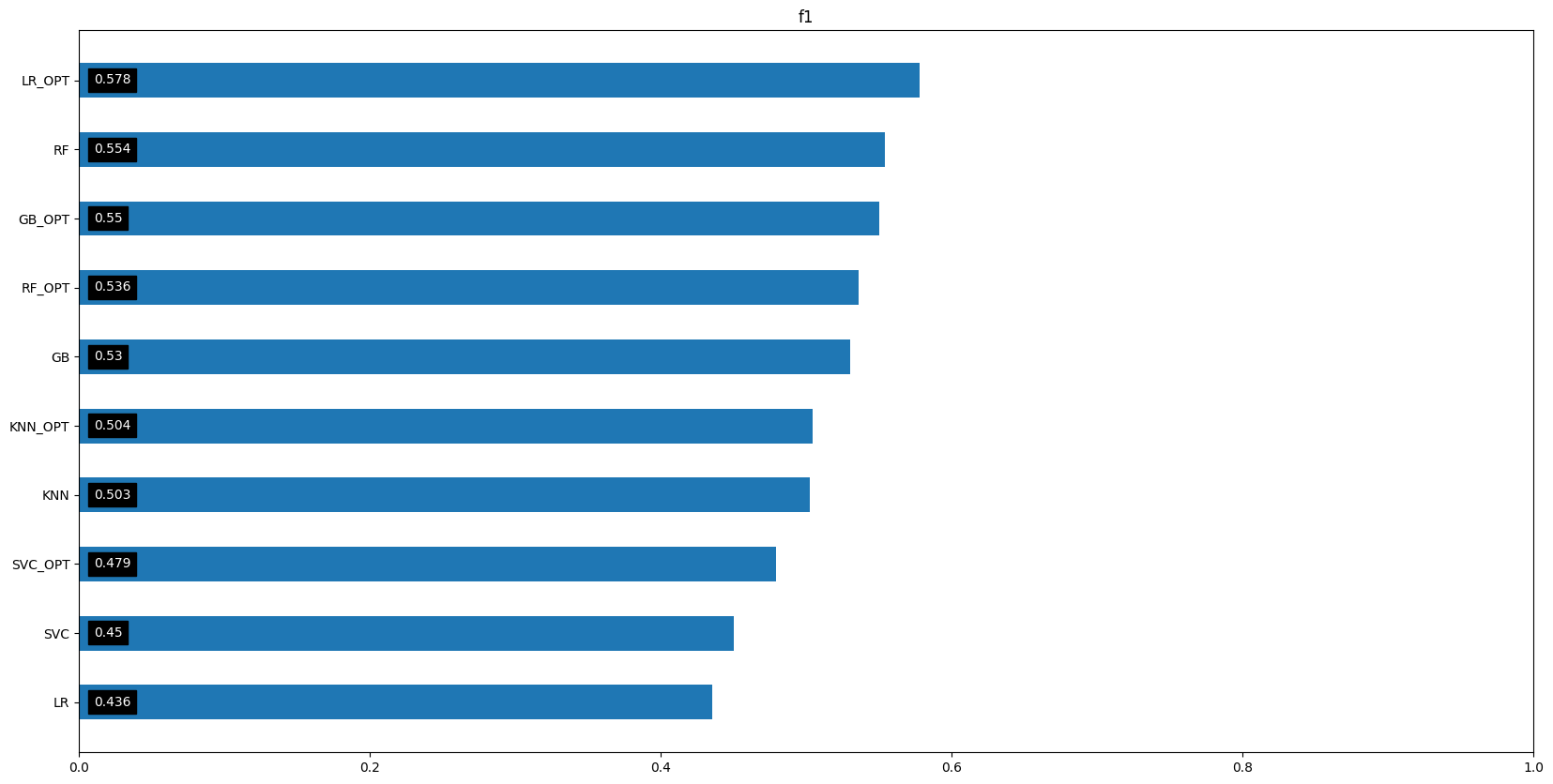


Рисунок 49 – Диаграмма значений метрики f1

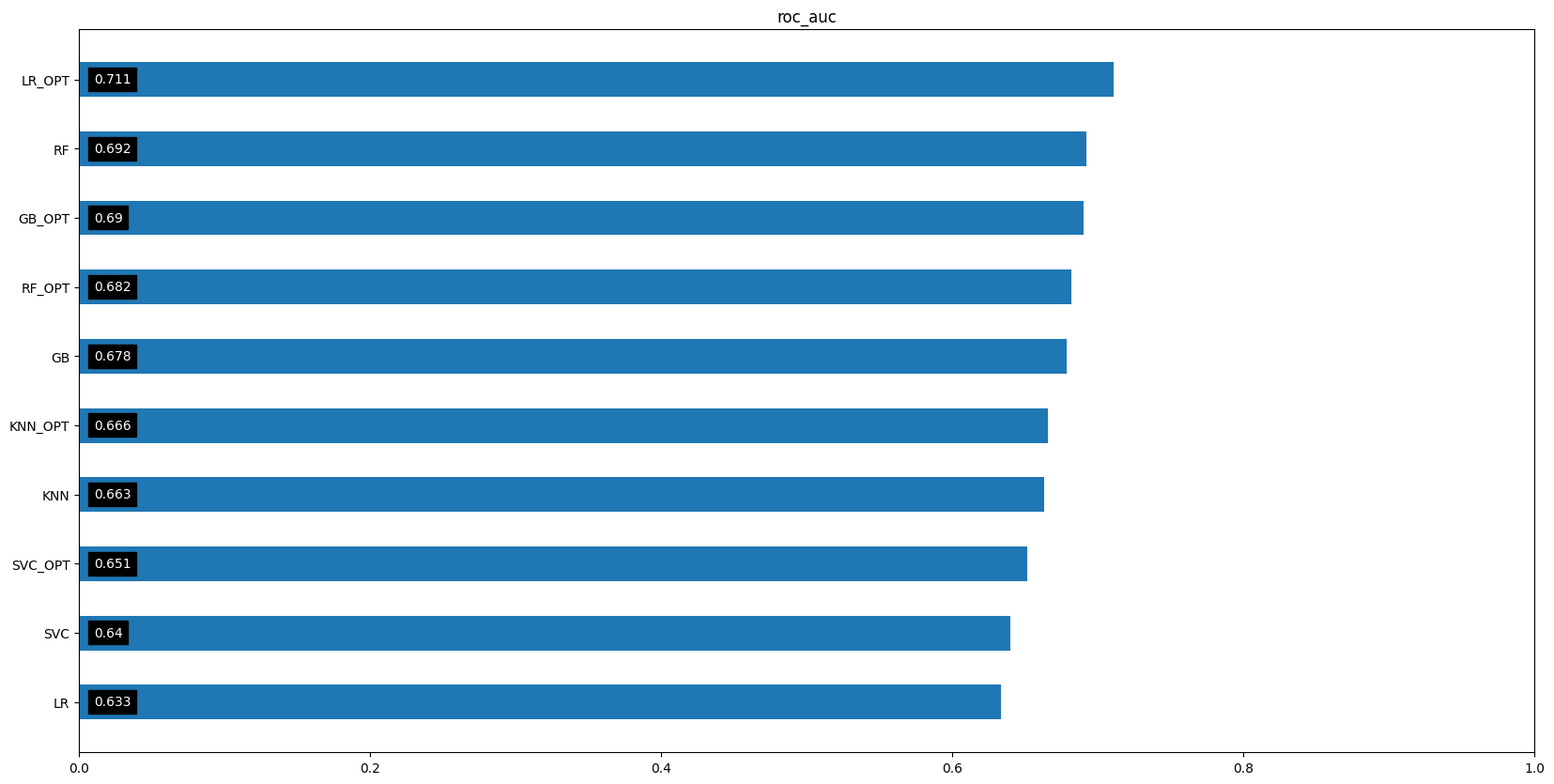


Рисунок 50 – Диаграмма значений метрики roc\_auc

**Выводы:**

* Оптимизация **метода ближайших соседей** привела к несколько *лучшим* результатам по всем метрикам
* Оптимизация **логистической регрессии** привела к намного *лучшим* результатам по всем метрикам
* Оптимизация **метода опорных векторов** привела к несколько *лучшим* результатам по всем метрикам
* Оптимизация **случайного леса привела** к несколько *худшим* результатам по всем метрикам
* Оптимизация **градиентного бустинга** привела к несколько *лучшим* результатам по всем метрикам
* Лучшие результаты были показаны методом **логистической регрессии** после оптимизации

# **Создание веб-приложения**

В качестве модели была выбрана логистическая регрессия.

Сначала импортируем необходимые библиотеки, в функции load\_data загружаем набор данных, обрабатываем данные: удаляем дубликаты, строки с пропусками и лишние столбцы, заменяем столбцы аналогично тому, как это было сделано ранее, кодируем категориальные признаки с помощью LabelEncoder и масштабируем числовые с помощью MinMaxScaler.

Затем в фукнции preprocess\_data разделяем данные на тренировочную и тестовую выборку с помощью функции train\_test\_split.

Создаем функцию class\_accuracy\_score, которая будет вычислять значение accuracy для предсказаний модели для каждого класса по отдельности.

Устанавливаем с помощью st.sidebar.header заголовок боковой панели. С помощью st.sidebar.slider создаем ползунки для регулирования значений C и class\_weight, с помощью st.sidebar.radio создаем перключатель для значения solver.

С помощью st.checkbox создаем чекбокс для отображения матрицы корреляции набора данных.

Используя значения, полученные из переключателей и ползунков, создадим и натренируем модель.

С помощью st.header отобразим заголовок сайта.

С помощью st.subheader отобразим подзаголовок для таблицы точностей.

С помощью st.dataframe отобразим результаты использования функции class\_accuracy\_score.

С помощью st.subheader отобразим подзаголовок для матрицы ошибок.

С помощью st.pyplot и ConfusionMatrixDisplay.from\_predictions отобразим матрицу ошибок.

Создадим функцию draw\_roc\_curve для отрисовки ROC кривой по предсказаниям модели.

С помощью st.subheader отобразим подзаголовок для ROC кривой.

С помощью st.pyplot и draw\_roc\_curve отобразим ROC кривую.

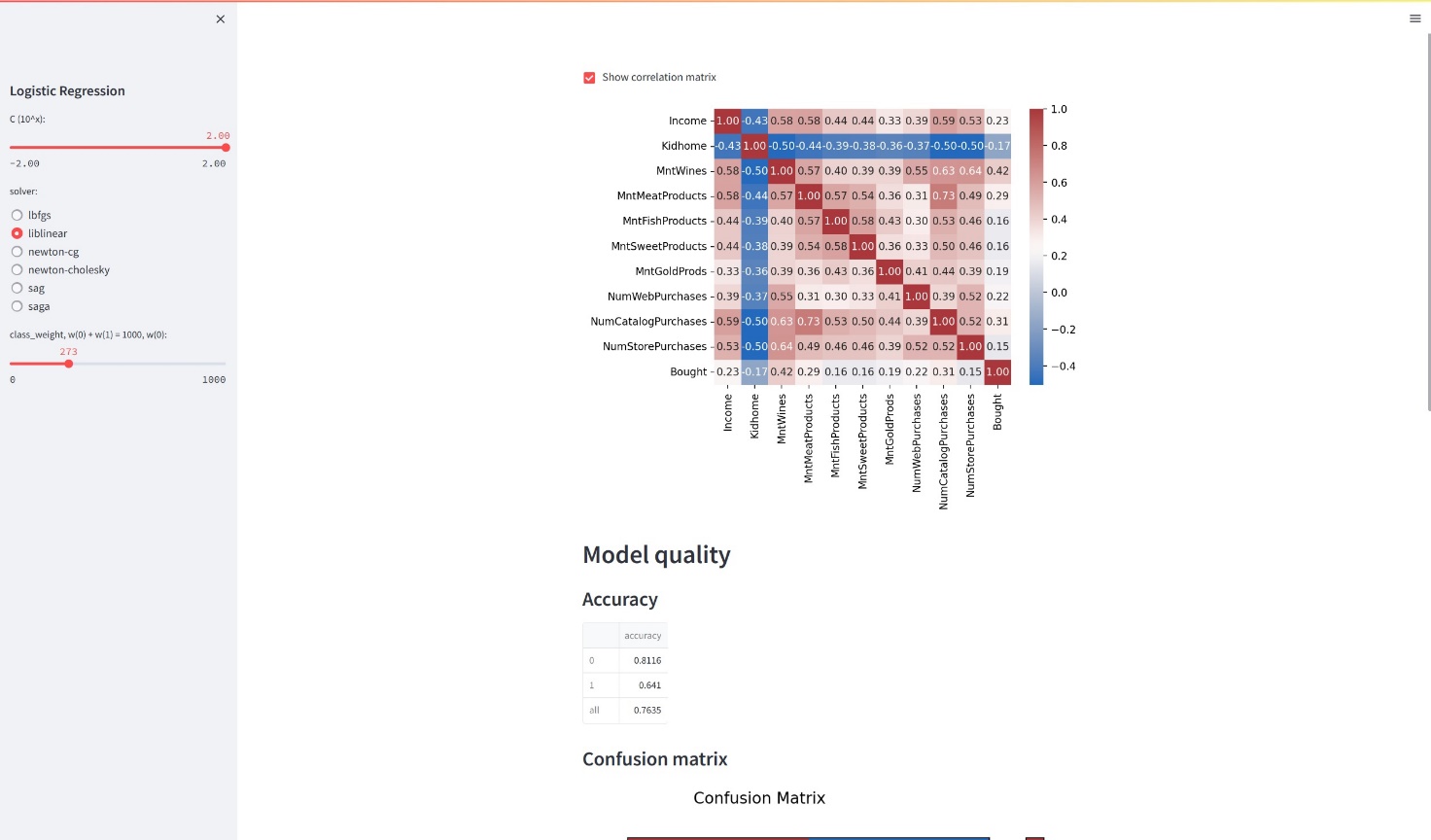


Рисунок 51 - Скриншот веб-приложения (1)

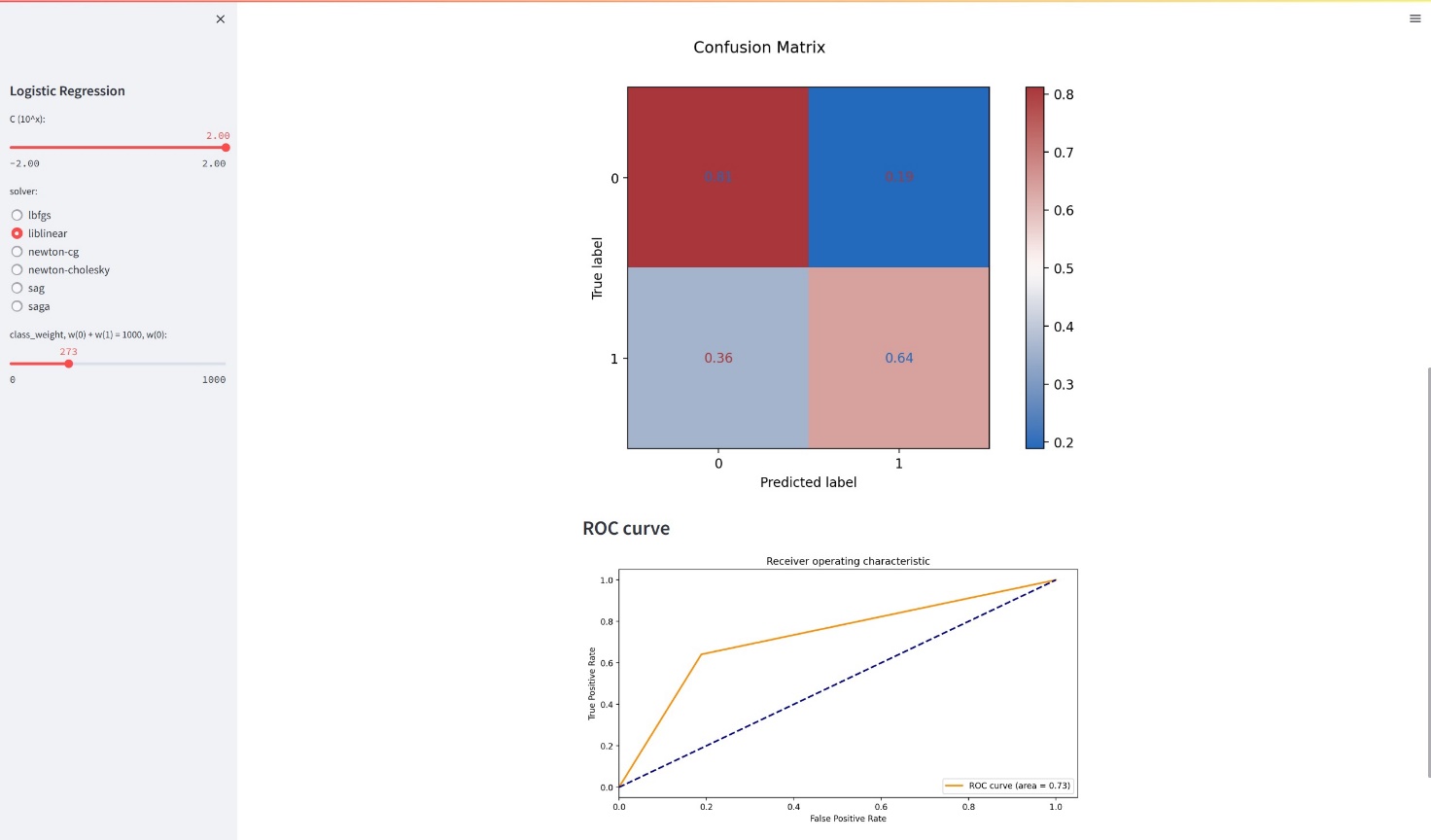


Рисунок 52 - Скриншот веб-приложения (2)

# **Заключение**

Оптимизация затрат на рекламу – важная задача для любого бизнеса. Анализ данных пользователей с помощью алгоритмов машинного обучения может помочь решить эту проблему.

В рамках НИР была рассмотрена задача классификации клиентов компаний с помощью методов машинного обучения. Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как метод ближайших соседей, логистическая регрессия, метод опорных векторов, случайный лес и градиентный бустинг.

В результате исследования было показано, что большинство использованных методов могут достичь приемлемых результатов при должном уровне оптимизации, однако решение задачи на данном этапе сводится скорее к сбору большего количества данных. В данном исследовании лучший потенциал показал метод логистической регрессии.

# **Список использованной литературы**

1. API Reference // Scikit Learn URL: https://scikit-learn.org/stable/modules/classes.html (дата обращения: 30.04.2023).
2. Balanced Accuracy: When Should You Use It? // MLOps blog URL: https://neptune.ai/blog/balanced-accuracy (дата обращения: 01.05.2023).
3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://scikit-learn.org/stable/tutorial/statistical\_inference/supervised\_learning.html (дата обращения: 25.04.2023).
4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses\_current/wiki/COURSE\_TMO\_SPRING\_2023/ (дата обращения: 25.04.2023).