Модель склонности клиента к приобретению покупки машиноместа

самолет

Работу выполнили студенты группы ИСП-22:

Кривобокова Ольга

Лесницкая Татьяна

Титова София

Преподаватель:

Коновалов Игорь Васильевич

самолет

О кейсе

Постановщик задачи:

Компания «Самолет»

Название кейсового задания:

Модель склонности клиента к приобретению машиноместа

Цель:

Разработать модель, позволяющую прогнозировать вероятность покупки клиентами дополнительных услуг в частности, приобретения машиномест в паркинге

Этапы работы

1

Предварительный анализ данных

2

Разделение задач и формирование индивидуальных решений

3

Предобработка данных

каждый участник

4

Создание и обучение моделей

каждый участник

5

Изучение и сравнение полученных решений

6

Объединение лучших подходов в одно решение

7

Обучение, настройка и оптимизация итоговой модели

8

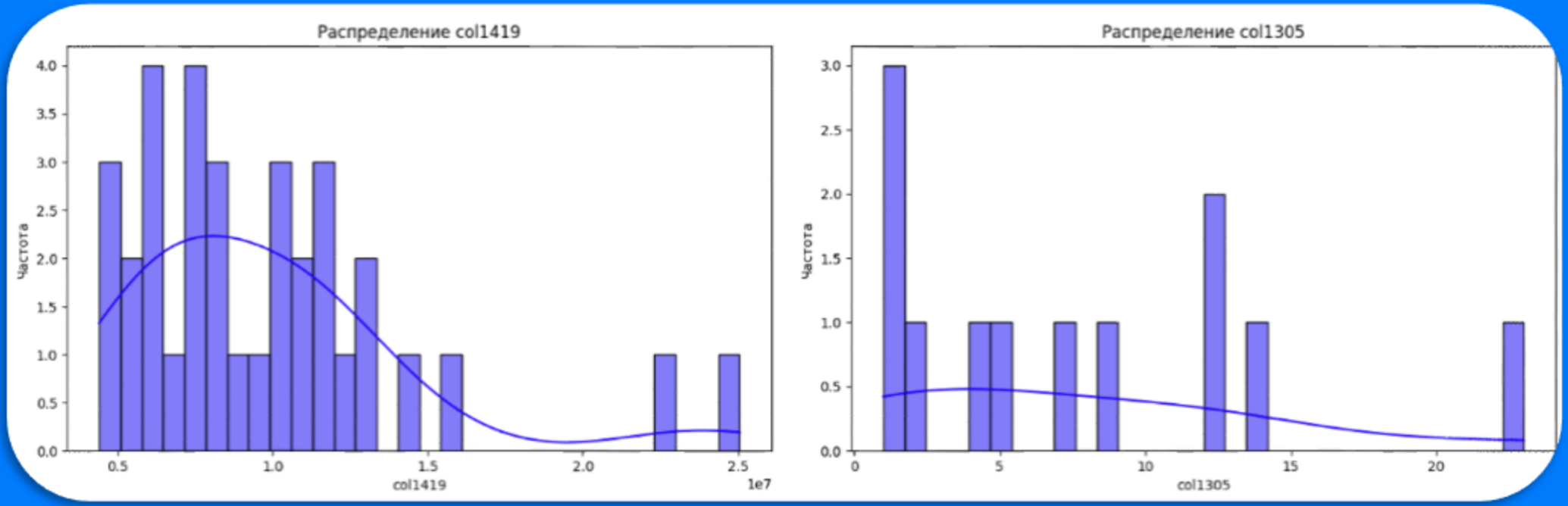
Получение и интерпретация итогового результата

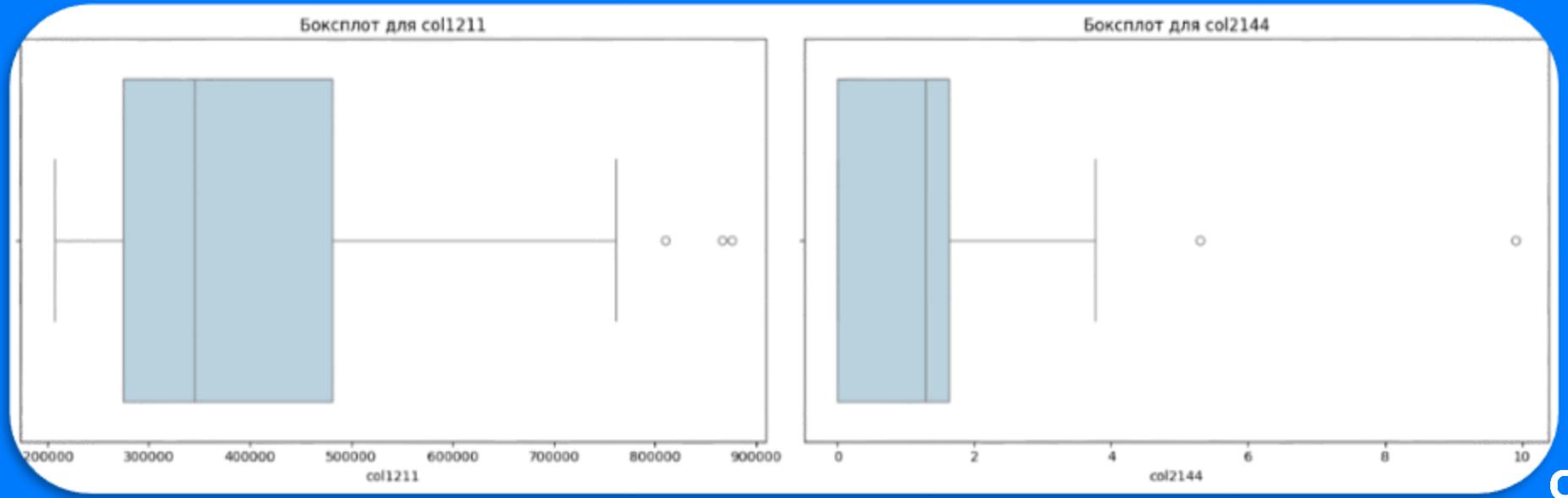
самолет

Предварительный анализ данных

Мы выяснили, что данные не подчиняются нормальному распределению, что объясняет выбор MinMaxScaler вместо StandardScaler, так как он лучше справляется с выбросами

Также был обнаружен сильный дисбаланс классов, что может привести к переобучению и ухудшению предсказаний для малочисленного класса

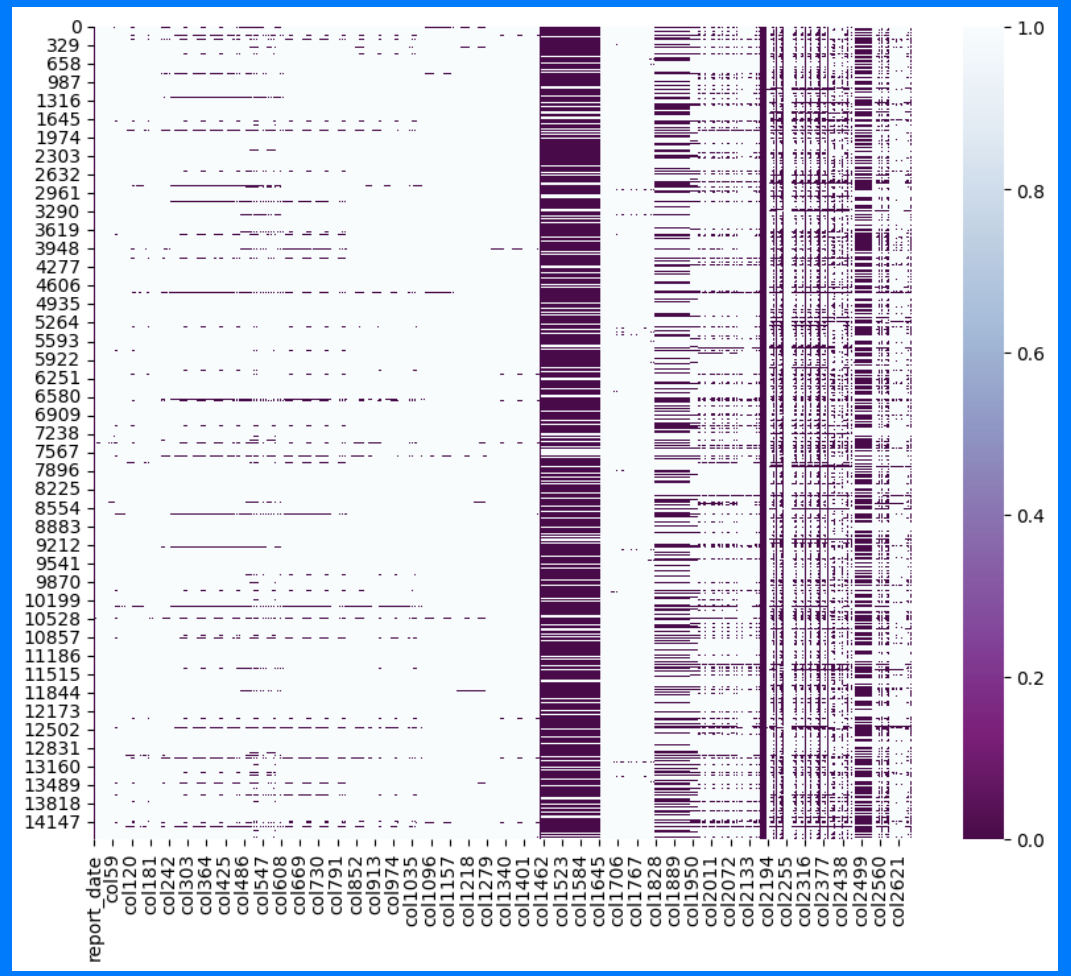

Распределение col1419
Частота
col1419
1e7
Распределение col1305
Частота
col1305


Боксплот для col1211
col1211
Боксплот для col2144
col2144

самолет

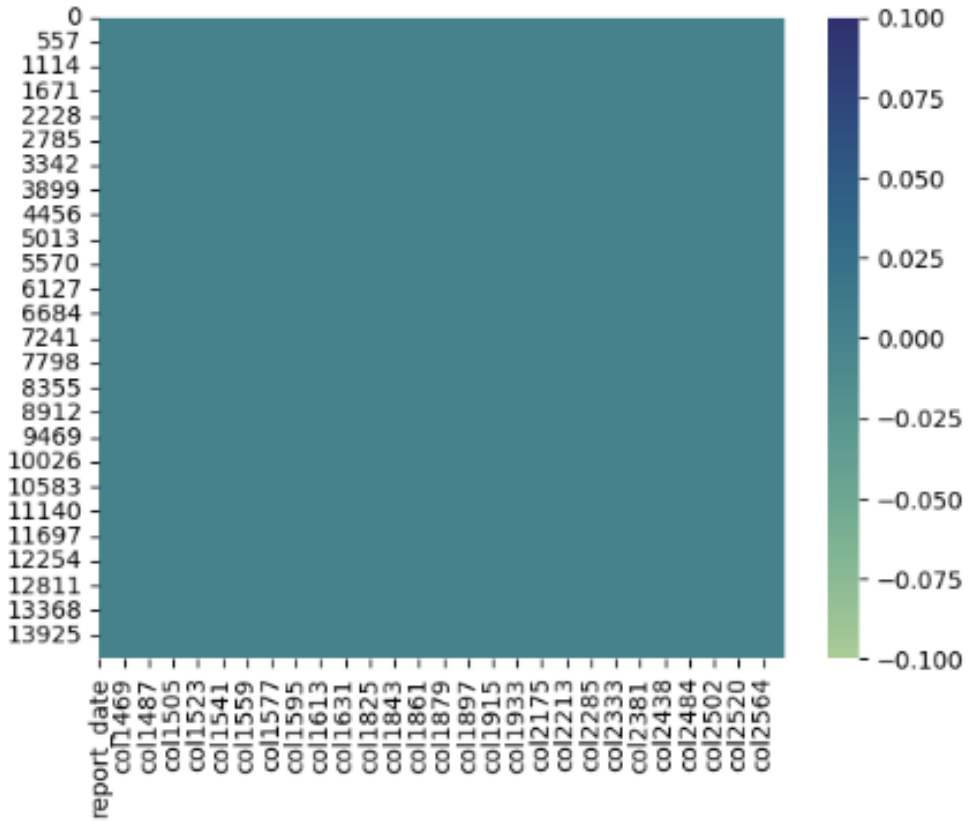
Обработка данных

- **Удалили колонки** с более чем **70% пропущенных значений.**
- Провели **поиск дубликатов** и удалили их для предотвращения искажения результатов.
- Искали и удаляли **мусорные данные**, такие как ссылки, некорректные значения.
- Поработали с **object-колонками.**



самолет

Вывод по обработке данных



Выбросы: выявлены с помощью IQR и Isolation Forest.

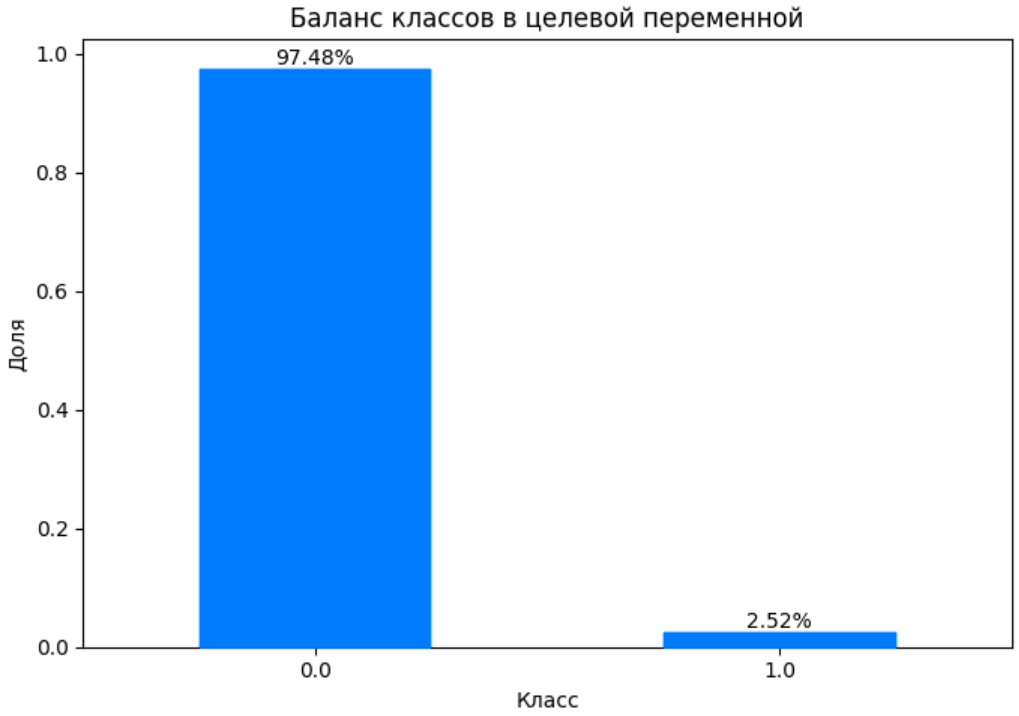
Пропуски: числовые признаки заполнены медианой, категориальные — значением 'most_frequent'.

Категориальные данные: преобразованы с помощью OrdinalEncoder.

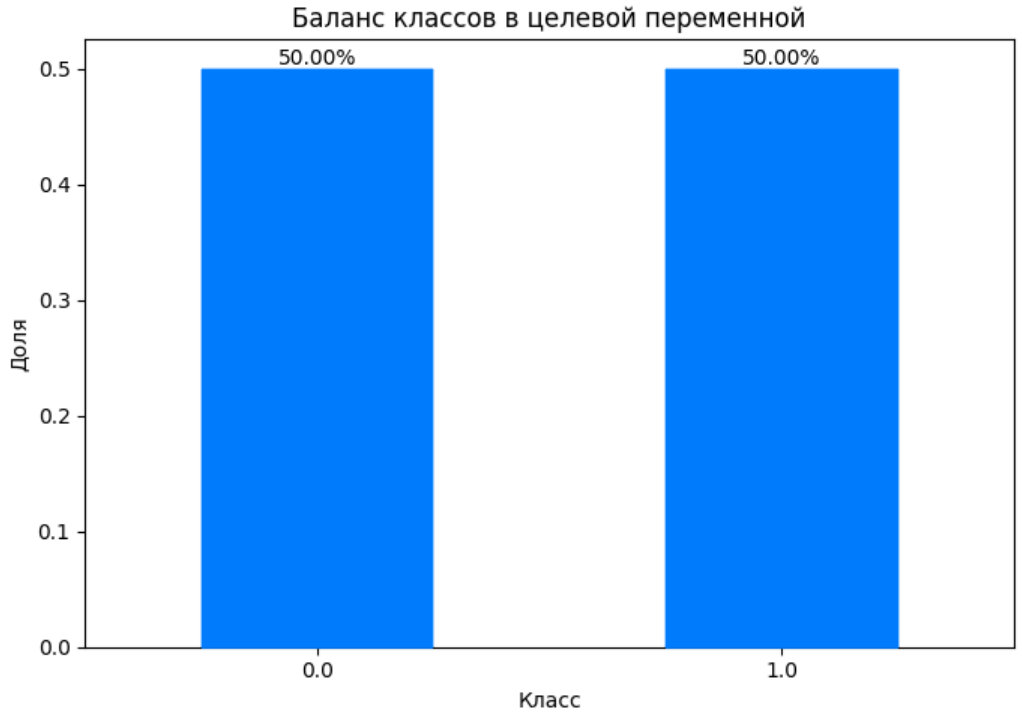
Масштабирование: использован MinMaxScaler.

Дисбаланс классов и его решение

При анализе целевой переменной был выявлен дисбаланс классов:
•Класс **0.0** составляет **97.5%**,
•Класс **1.0** — лишь **2.5%**.

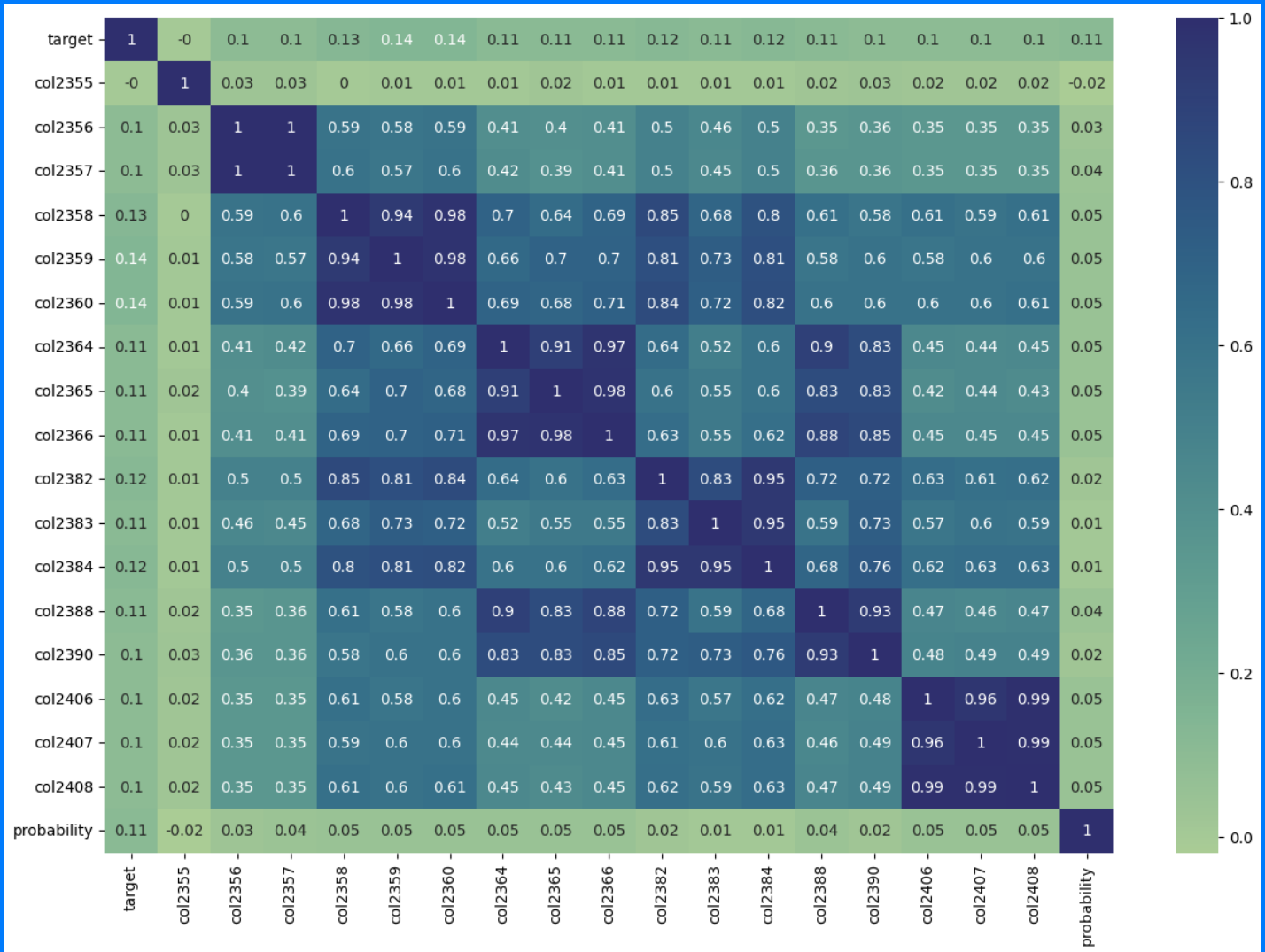


Решение проблемы дисбаланса:
Для устранения дисбаланса применён метод **SMOTETomek**, который позволяет сбалансировать классы
и улучшить качество модели.



самолет

Отбор признаков



Корреляционный анализ выявил признаки с влиянием на target ($|r| \geq 0.09$).

Избыточная корреляция: признаки с высокой зависимостью ($|r| \geq 0.75$) объединены путем усреднения, избыточные столбцы удалены.

самолет

Гиперпараметры

Гиперпараметры модели подбирались с помощью **Optuna.** Были оптимизированы следующие параметры:

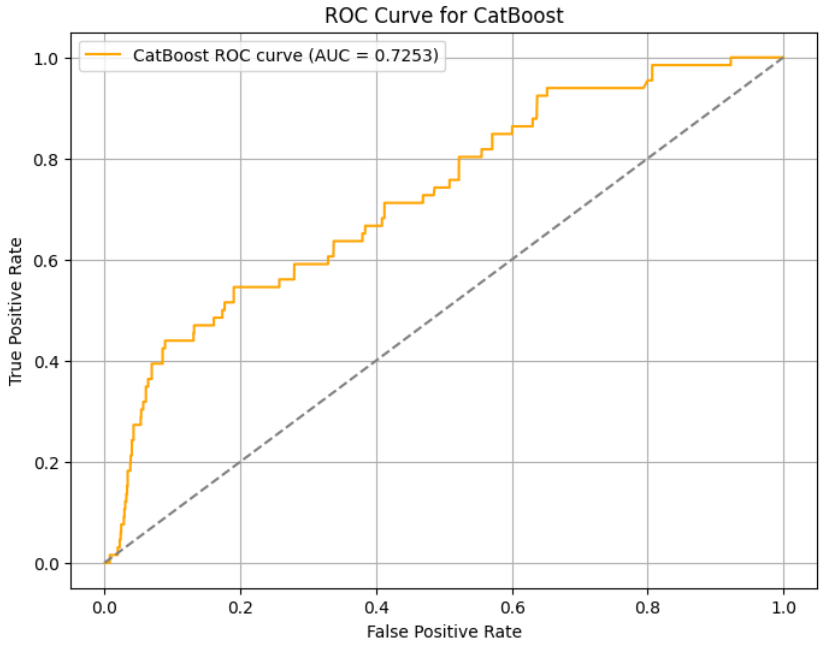
- количество деревьев
- скорость обучения
- доля обучающих данных для дерева
- доля признаков
- максимальная глубина
- L1 и L2 регуляризация

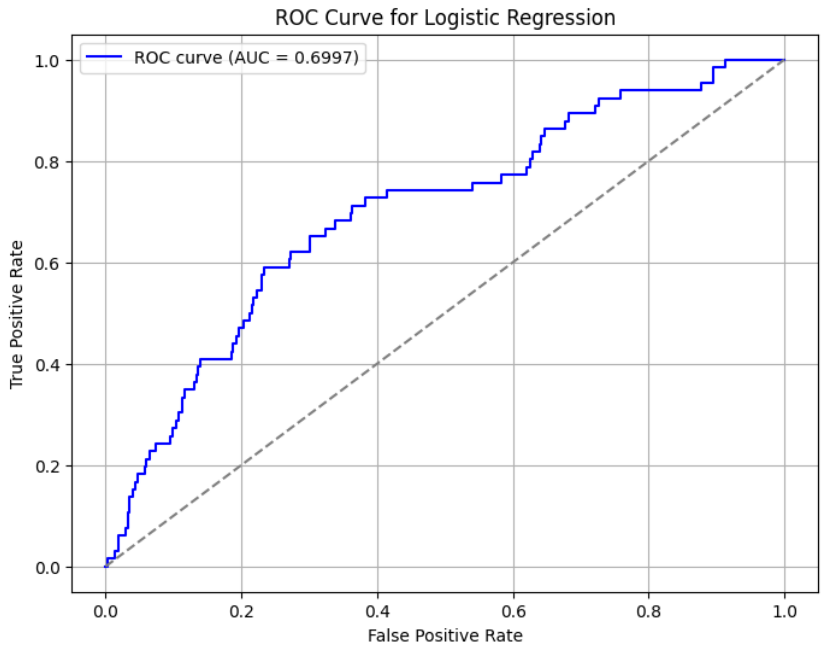
Модели

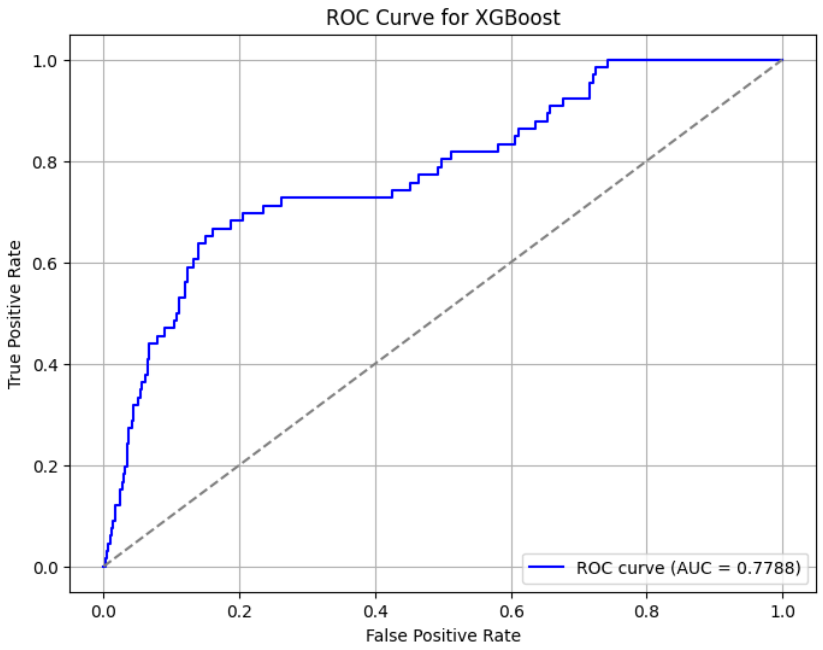
Catboost

Logistic regression

Xgboost







Выбор модели Xgboost основан на хороших результатах, которые она показывала на сырых данных по сравнению с моделями CatBoost и Logistic Regresion

самолет

Проблемы с которыми мы столкнулись

- Дисбаланс классов
- Переобучение
- Разреженность данных

Вывод

У нас получилось собрать лучшие на наш взгляд способы, алгоритмы и модели машинного обучения для решения поставленных задач из тех, которые мы попробовали.