Отчет по работе 2го интенсива

Работу выполнил: Хусаинов Марат

самолет

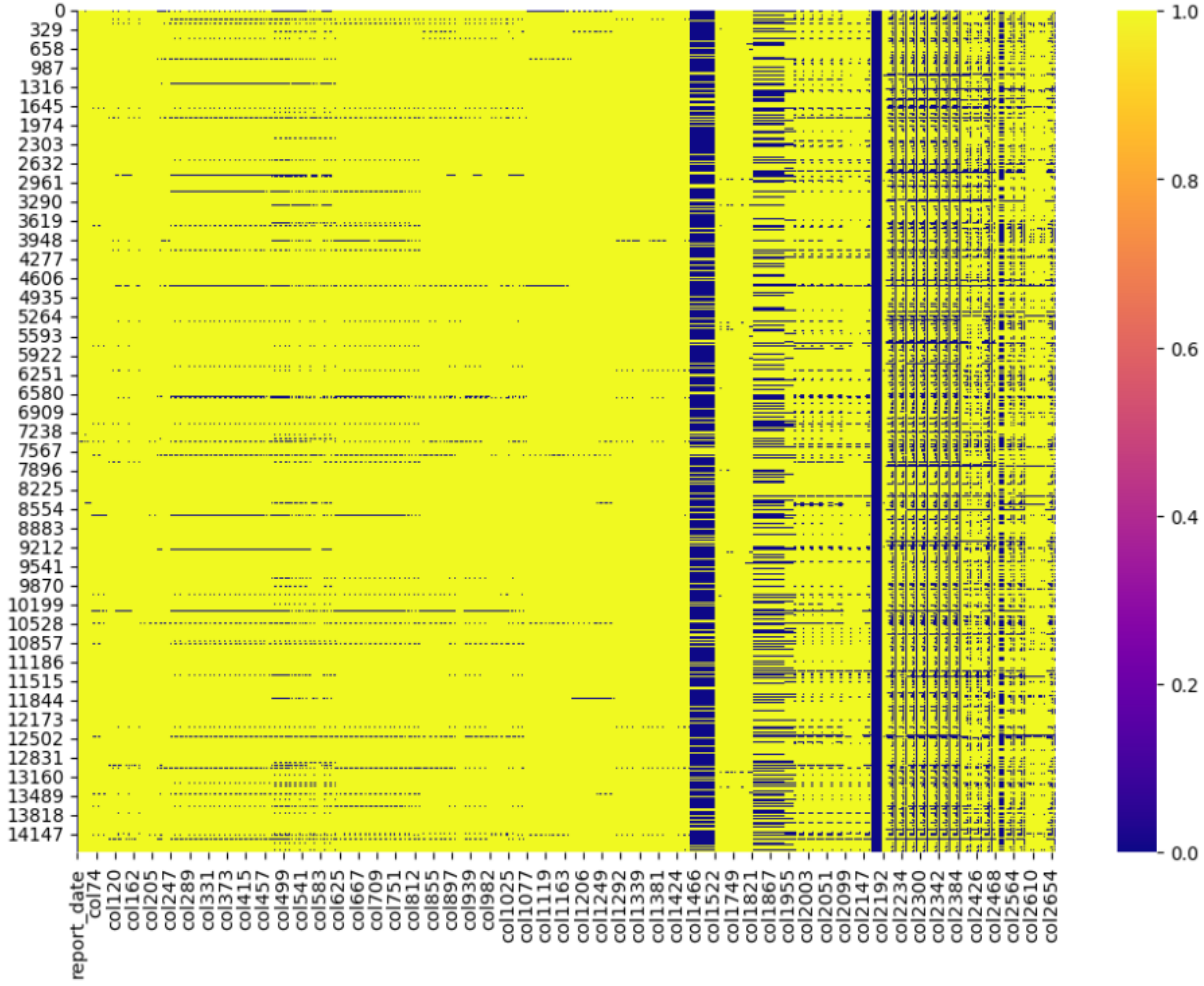
О кейсе

Название : Модель склонности клиента к приобретению машиноместа

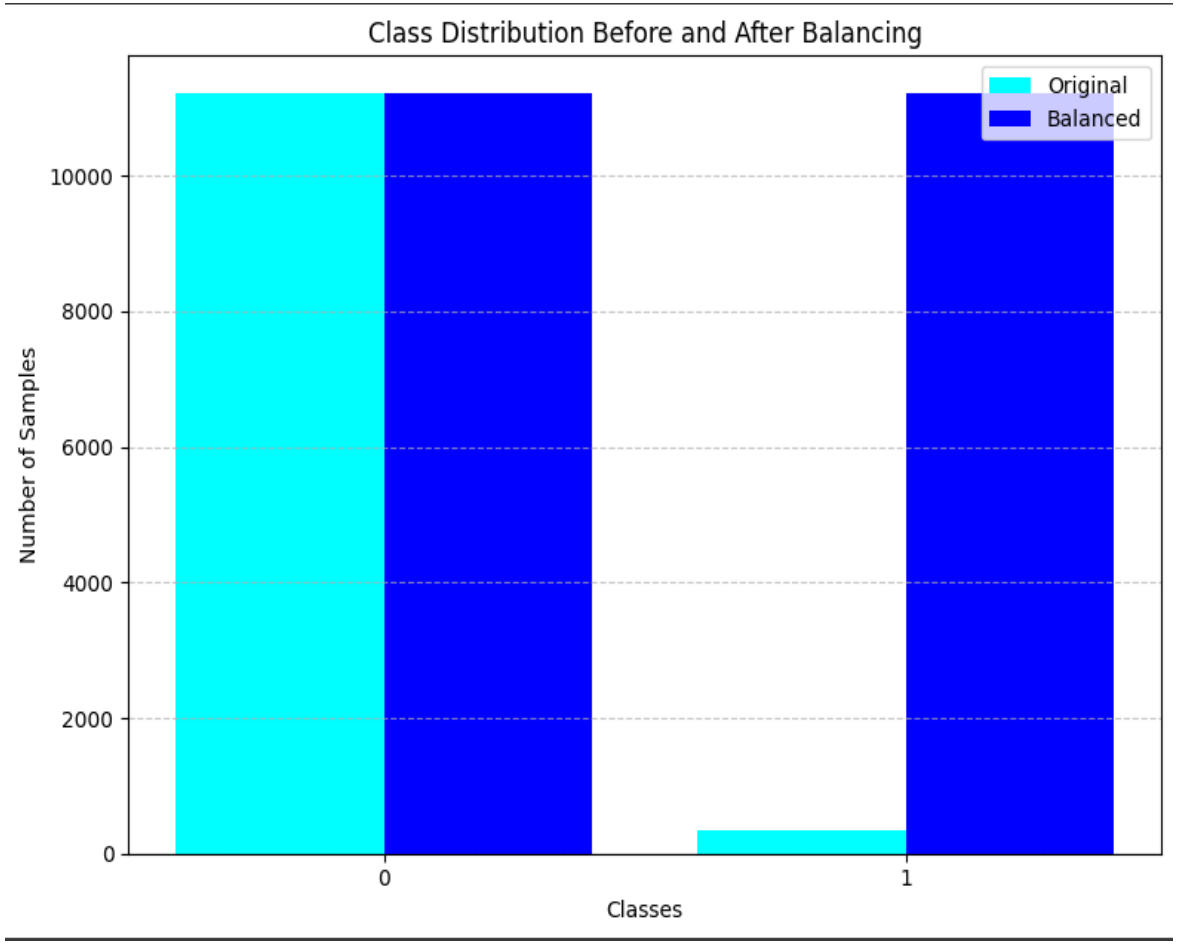
Описание задания: На основе больших данных о предыдущем опыте взаимодействия с клиентами разработать модель, позволяющую прогнозировать вероятность покупки клиентами дополнительных услуг в частности, приобретения машиномест в паркинге. Среди клиентов компании - владельцев квартир необходимо выделить покупателей, наиболее склонных к покупке машиноместа. С такими клиентами будет проводиться коммуникация (смс, эл. письмо) с предложением приобрести машиноместо.

Графики о датасете

Тепловая карта изначально



Распредление классов



самолет

Модели

XGBClassifier — это реализация алгоритма **градиентного бустинга** (Gradient Boosting) с использованием деревьев решений. Он является одной из самых популярных и мощных моделей машинного обучения благодаря своей высокой производительности, гибкости и эффективности.

RandomForestClassifier — это алгоритм машинного обучения, который относится к семейству **ансамблевых методов** (ensemble methods). Он строит ансамбль деревьев решений и объединяет их предсказания для повышения точности и устойчивости модели.

Метрики в числах

Metrics for RandomForestClassifier:

- Accuracy: 0.9820
- Precision: 0.8667
- Recall: 0.7803
- F1-Score: 0.8173
- AUC: 0.97

Metrics for XGBClassifier:

- Accuracy: 0.9824
- Precision: 0.8284
- Recall: 0.8857
- F1-Score: 0.8546
- AUC: 0.97

Обработка данных

Удаление колонок заполненных меньше чем на 50%

Заполнение нулями NaN значений.

Заполнение пустыми строками NaN значений в нечисловых колонках

Удаление специфических данных:

Hash - непонятные символы

Ссылки - ненужные данные

Удаление повторяющихся данных (дубликатов)

самолет

Предобработка данных для обучения

1

Баласировка данных.

Поскольку положительный класс встречался куда реже отрицательного использовалась модель для балансировки SMOTE

2

Кодировка категориальных принципов с помощью LabelEncoder

3

Попытка обработки выбросов моделью IsolationForest

(не получилось)

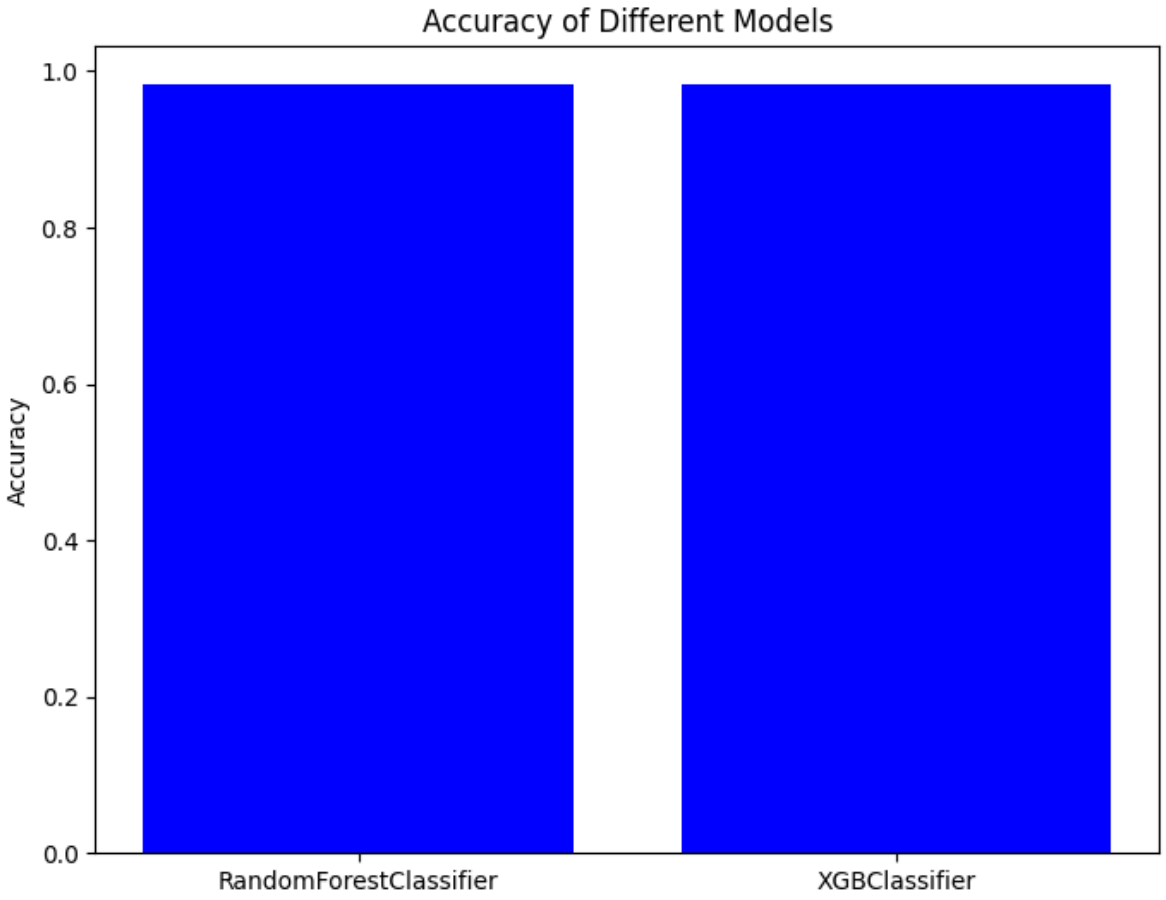
4

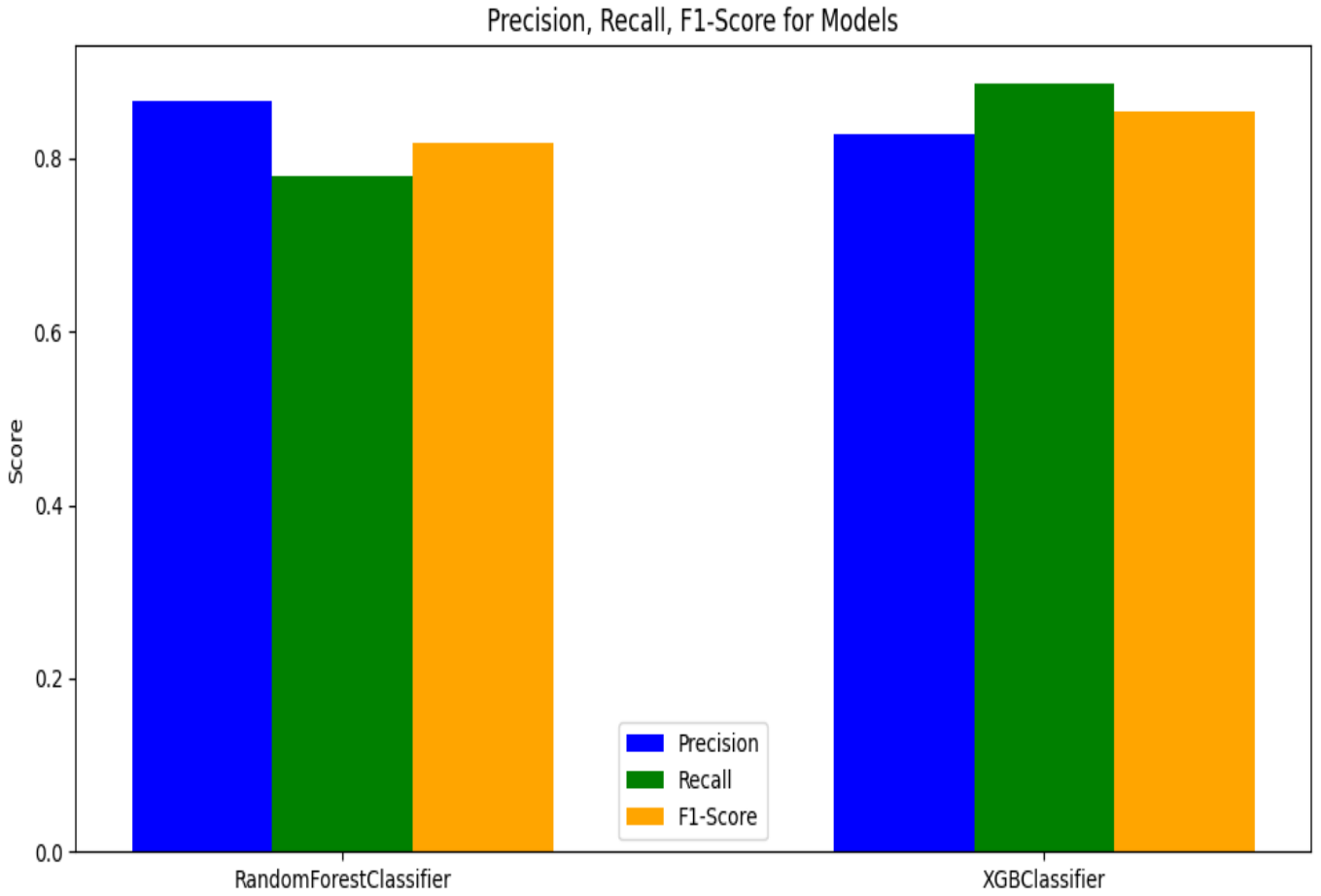
Разделение на тестовую и обучающую выборки

самолет

Результаты

Обе модели были обучены, для них были подобраны гиперпараметры с помощью GridSearchSv



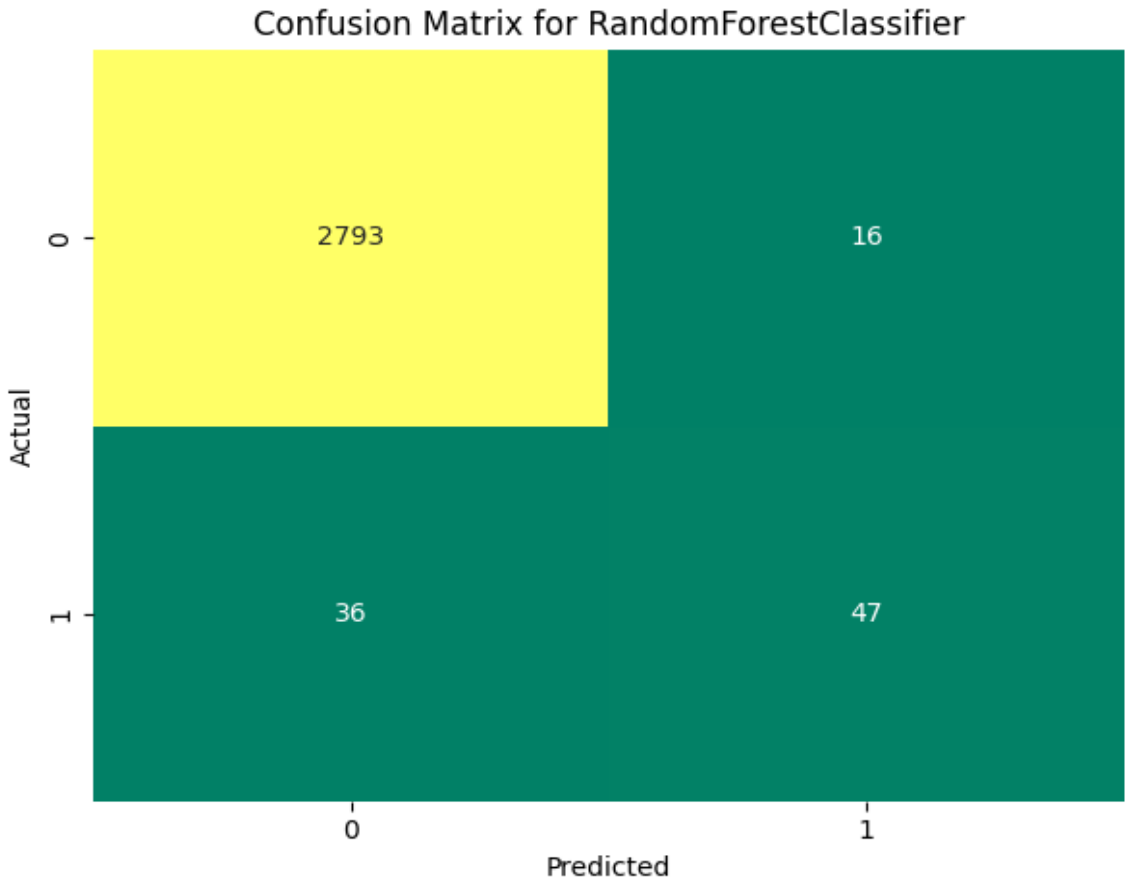


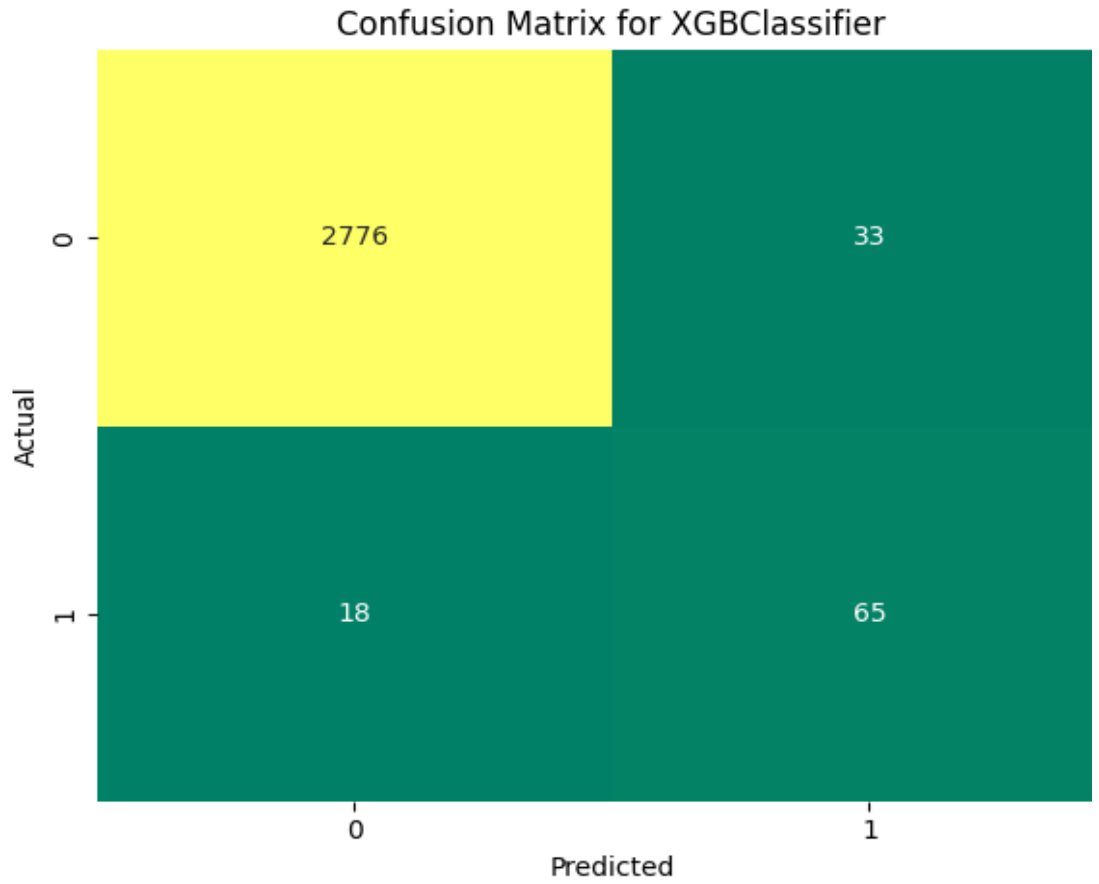
Графики разных метрик

самолет

Выводы по тестовой выборке:

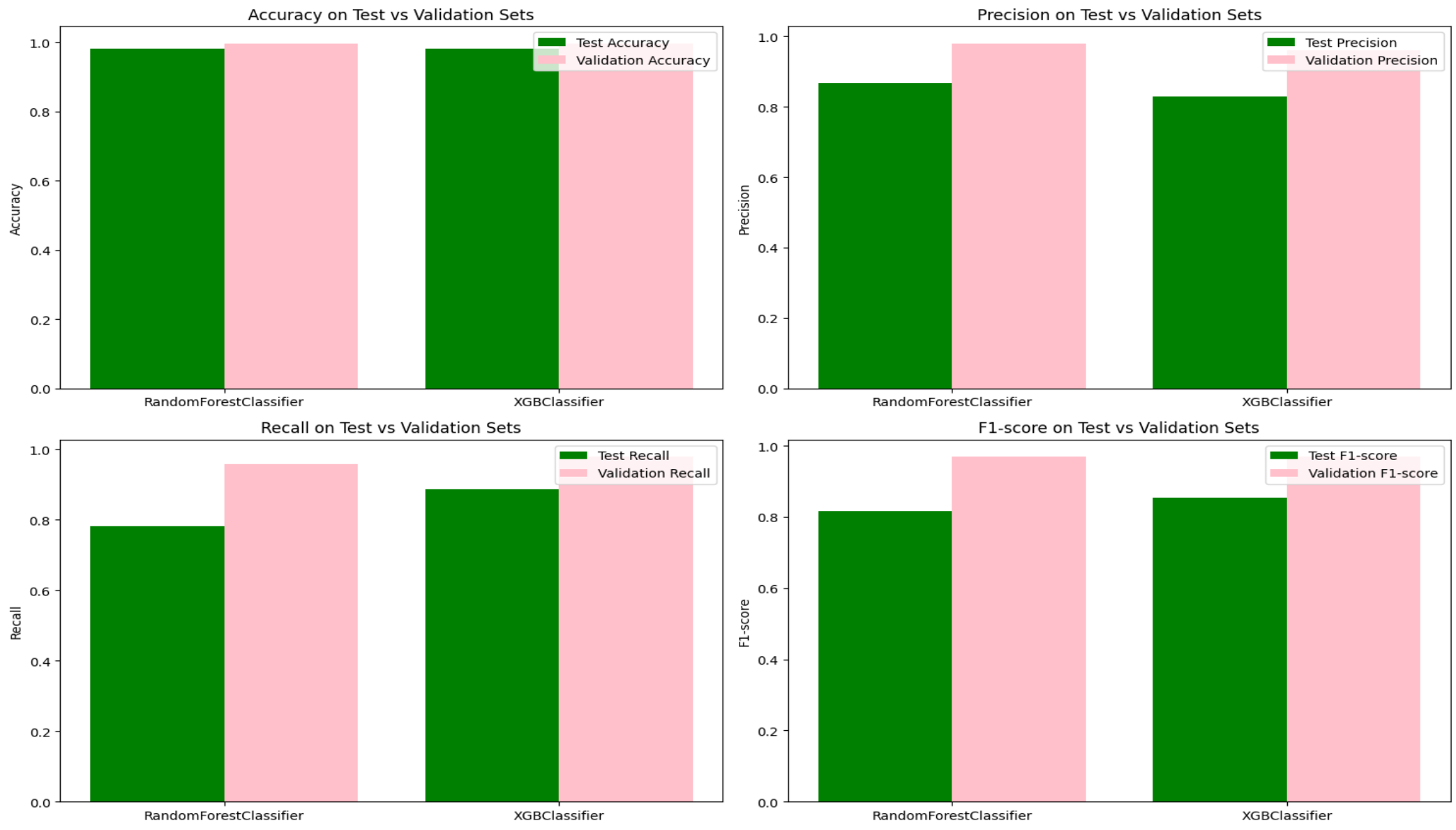
Отличия между моделями в производительности не существенны





самолет

Сравнение результатов на валидационном датесете с тестовым



самолет

Цифры метрик валидации

Validation Metrics for RandomForestClassifier:

Accuracy: 0.9964

Precision: 0.9795

Recall: 0.9581

F1-Score: 0.9686

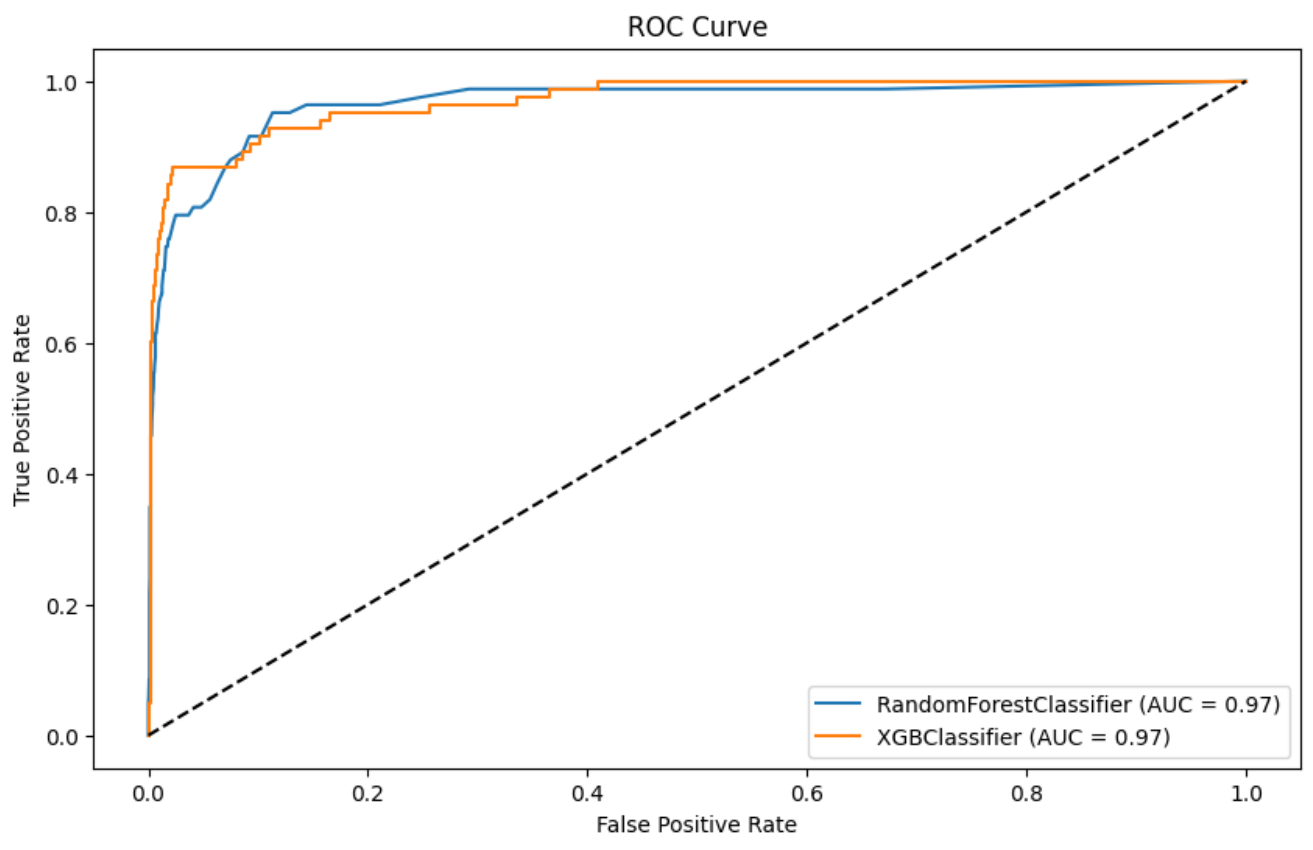
Validation Metrics for XGBClassifier:
Accuracy: 0.9963

- **Precision: 0.9597**
- **Recall: 0.9781**
- **F1-Score: 0.9687**

самолет

Вывод:

Я считаю что задание выполнено, обе модели показали хорошие результаты на тестовом и валидационном датасетах. Модель на основе Gradient Boosting показала себя лучше. В валидационном датесете она совершила 48 ошибок, против 52 ошибок у RandomForest



самолет