# Модель склонности клиента к приобретению машино-места

самолет

# Цель

На основе больших данных о предыдущем опыте взаимодействия с клиентам поиска клиентов, наиболее склонных к приобретению машино-места.

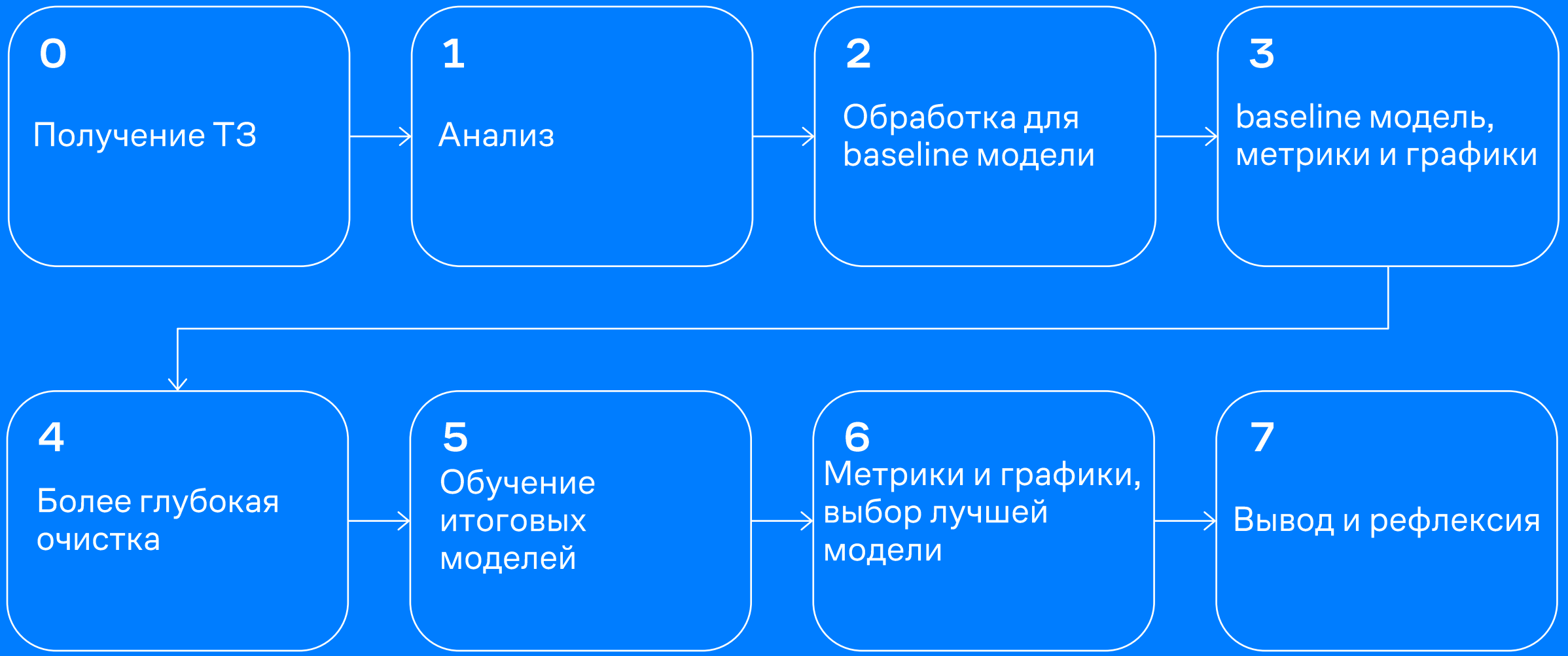
Разработать модель, позволяющую прогнозировать вероятность покупки клиентами дополнительных услуг в частности, приобретения машино-мест в паркинге.

Среди клиентов компании - владельцев квартир необходимо выделить покупателей, наиболее склонных к покупке машино-места. С такими клиентами будет проводиться коммуникация с предложением приобрести машино-место.

# Инструменты

- Vscode - как редактор кода
- Python - как язык программирования
- Pandas - для обработки данных
- Sklearn - для обучения моделей
- Matplotlib - для вывода графиков

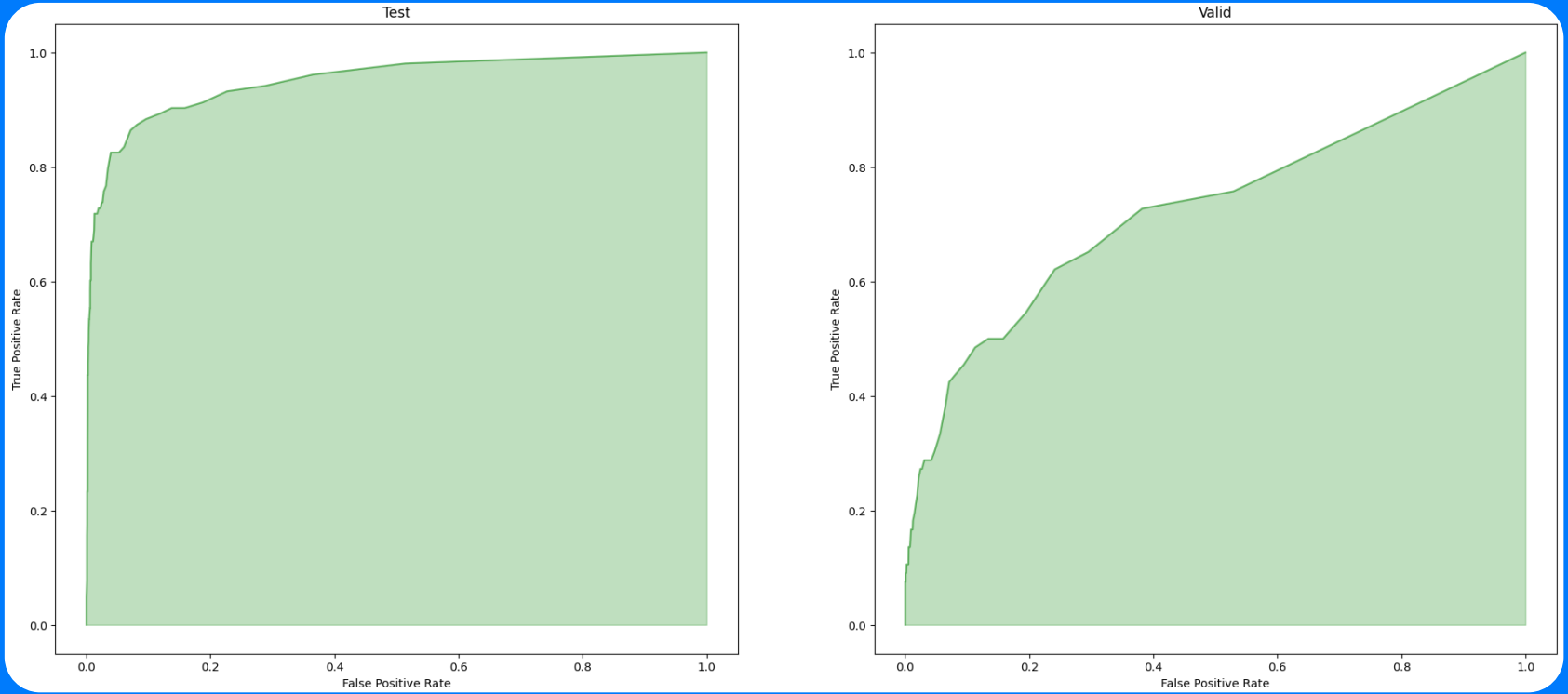
# Ход работы


0
Получение ТЗ
1
Анализ
2
Обработка для baseline модели
3
baseline модель, метрики и графики
4
Более глубокая очистка
5
Обучение итоговых моделей
6
Метрики и графики, выбор лучшей модели
7
Вывод и рефлексия

# Baseline модель

Для простой модели, которая будет использоваться для сравнения, можно крайне поверхностно почистить данные, например заменить все пропуски на ноль, а объектные данные – удалить.

У нас вышла оценка 0.95 по метрике ROC-AUC на тестовой части и 0.72 на валидационном датасете, что не очень хорошо, но это и понятно, на то она и baseline.




самолет

# Работа с данными

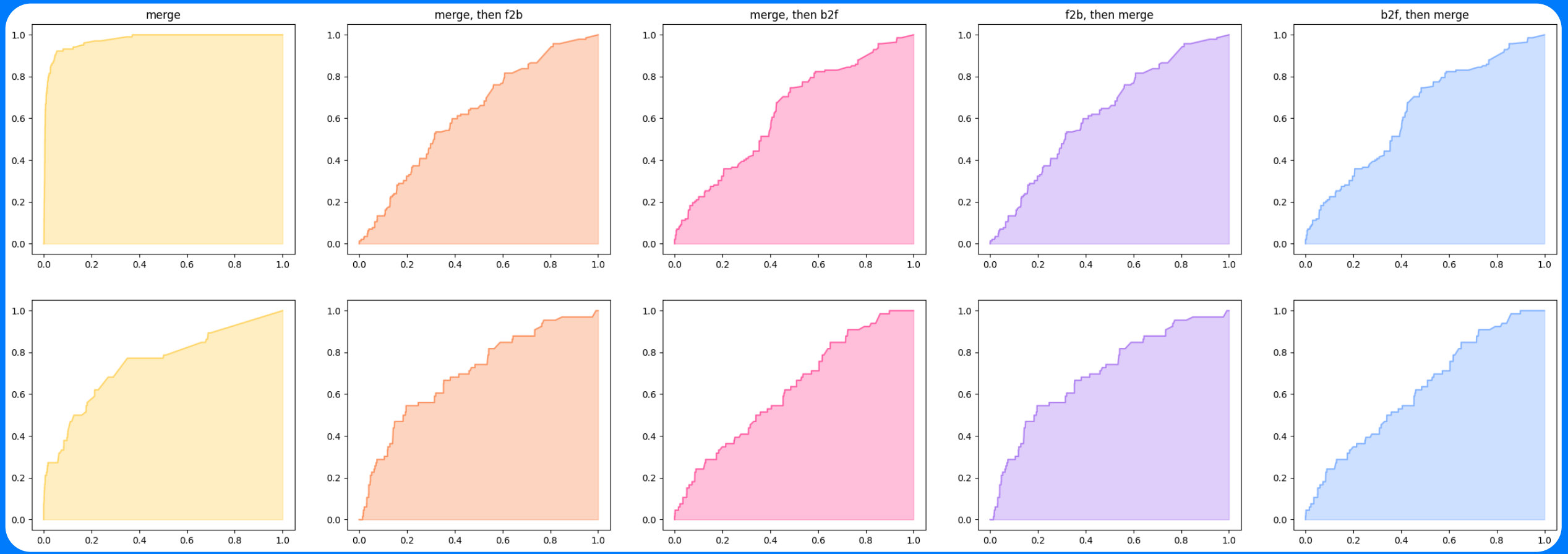
Из-за большого количества колонок и пропусков было решено совместить колонки с похожими значениями и, так как у нас есть записи клиентов с разных дат, заполнить пропуски уже известными значениями, только оставались вопросы:

- В каком порядке заполнять пропущенные данные: из новых в старые или наоборот?
- Что лучше первым сделать, совместить колонки или заполнить пропуски?

Так было решено попробовать всё и сразу.

# Обучение моделей

После создания пяти разных датасетов, обучения модели на них и оценки их с помощью roc-auc, мы получили вот такую картину:




самолет

# Выбор модель

По результатам, слева направо(test-valid):

1. 0.97-0.75
2. 0.62-0.70
3. 0.63-0.62
4. 0.58-0.62
5. 0.60-0.57

Из этот следует, что лучше всего себя показали первая и вторая модели, но первая, судя по метрике переобучилась, что оставляет вторую модель как наилучшую.

самолет

# Итог

Потенциально можно улучшить:

1. Переписать код, он крайне запутан
2. Посмотреть на другие модели, несмотря на то, что random forest classifier подходит почти идеально
3. Улучшить подход к обработке данных, ведь если у клиента нет ни одной записи по какой-либо колонке, она вся остаётся пустой

самолет

# Спасибо за внимание

Работу выполнил студент Артём Новасельский
группы ИСП-21

самолет

# Спасибо за внимание!

Работу выполнил студент Артём Новасельский группы ИСП-21

самолет